EdWorkingPaper No. 19-66

Effects of School Turnaround on K-3 Student Achievement

Gary T. Henry
Vanderbilt University

Shelby M. McNeill
Vanderbilt University

Erica Harbatkin
Vanderbilt University

This article contributes to the literature on school turnaround by examining the effect of the North Carolina Transformation (NCT) initiative, which was implemented in 75 low-performing schools after the state's efforts to turn around the lowest performing schools under Race to the Top ended, on student reading score growth in grades K-3. Reading score growth is measured using the mCLASS Dynamic Indicators of Basic Early Literacy (DIBELS) assessment. Utilizing a regression discontinuity design, we find that the NCT intervention had null effects on K-3 reading score growth across both the 2016 and 2017 school years.

VERSION: May 2019

Effects of school turnaround on K-3 student achievement

Gary T. Henry

Shelby M. McNeill

Erica Harbatkin

Vanderbilt University

Draft - please do not cite or distribute without permission of the authors

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

Improving student outcomes in chronically low-performing schools has been a central focus of policy and research throughout the twenty-first century. Beginning with *No Child Left Behind* (NCLB) and continuing on through Race to the Top (RttT), School Improvement Grants (SIG), the NCLB waivers, and now the Every Student Succeeds Act (ESSA), the federal government has tried to encourage states and districts to improve performance in low-performing schools through means such as adopting state accountability policies and implementing federally approved "turnaround" reforms. Research examining the effects of school turnaround under RttT, SIG, and the NCLB waivers has recently begun to emerge and has generally found mixed results in regards to student outcomes in grades 3 and above (Bonilla & Dee, 2017; Carlson & Lavertu, 2018; Dee, 2012; Dickey-Griffith, 2013; Dougherty & Weiner, 2017; Hemelt & Jacob, 2017; Henry & Harbatkin, 2018; Papay & Hannon, 2018; Sun, Penner, & Loeb, 2017; Zimmer, Henry, & Kho, 2017). However, no studies as far as we are aware have analyzed the impact of school turnaround efforts on student achievement in early elementary grades.

In this article, we contribute to the literature on school turnaround by focusing on early-grade student achievement in low-performing turnaround schools. We specifically estimate the overall effect of the North Carolina Transformation (NCT) initiative, which was implemented in 75 low-performing schools after the state's efforts to turn around the lowest performing schools under RttT ended, on student reading score growth in grades K-3 using a regression discontinuity design. Reading growth is measured using the mCLASS Dynamic Indicators of Basic Early Literacy (DIBELS) assessment. This analysis extends upon the work of Henry and Harbatkin (2018), which found that the NCT initiative had a negative effect on student test score change in grades 3 and above. Given the importance of developing reading skills by the third grade, this study focuses on a critical period for students' development.

This paper proceeds as follows. In the next section, we review studies of school turnaround interventions that have occurred since RttT and SIG began. We then describe the NCT intervention and the methods of our study before turning to a presentation of the results and associated validity checks. We end with a discussion of the implications for school turnaround policy and for future research.

Background on School Turnaround

Race to the Top, School Improvement Grants, and NCLB waivers

Under the Obama-era policies of Race to the Top (RttT) and School Improvement Grants (SIG), the federal government tried to disrupt the status-quo in the lowest performing 5 percent of schools through linking school reform funding with the required implementation of one of four turnaround models:

- Transformation: requires replacement of principal; schools can choose among a set of promoted practices aimed at school improvement
- Turnaround: requires replacement of principal and at least 50 percent of school staff; schools can choose among a set of promoted practices aimed at school improvement
- Restart model: requires replacement of local school management with other management organizations (i.e. charter)
- Closure: requires closure of the school and enrollment of the students in higher performing schools.

Of the four models, turnaround and transformation were the least extreme. Both models required the replacement of the school principal and implementation of promoted practices aimed at school improvement, such as extending the school day or implementing instructional reform. The restart and closure models required more dramatic interruption of the status quo through replacement of school management or closure of the school.

In contrast to school reform under RttT and SIG, school turnaround under NCLB waivers was a more flexible but resource-limited approach. Specifically, interventions under waiver-based reforms were supported through existing Title 1 funds rather than additional federal funding. They also did not require schools to implement any federally promoted intervention models. Lastly, interventions under NCLB waivers largely targeted schools with the largest achievement gaps or lowest achievement among traditionally underserved student subgroups rather than the lowest performing 5 percent of schools.

Studies of School Turnaround

Research examining the effects of school turnaround under RttT, SIG, and the NCLB waivers has recently begun to emerge over the last five years as more districts and states have implemented turnaround strategies. Summaries of these turnaround studies that highlight the specific location of the intervention, grade levels included in analyses, and impact on student achievement are presented in Table 1.

Table 1 ABOUT HERE

Overall, these studies of school turnaround have found mixed results. Of the ten studies highlighted in Table 1, five found only positive effects of school turnaround on student achievement (Bonilla & Dee, 2017; Carlson & Lavertu, 2018; Papay & Hannon, 2018; Schueler, Goodman, & Deming, 2017; Sun, Penner, & Loeb, 2017). For example, Papay and Hannon (2018) found that school turnaround efforts focused on the 35 lowest performing schools in Massachusetts yielded significant positive effects in math and English Language Arts (ELA) achievement for students in those schools. These effects emerged in the first year of turnaround and grew through the fourth year (Papay & Hannon, 2018). Studies of school turnaround in Tennessee (Zimmer, Henry, & Kho, 2017) and in Los Angeles Unified School District (LAUSD) (Strunk, Marsh, Hashim, Bush-Mecenas, & Weinstein, 2016) found both positive and negative or null effects depending on the school reform model implemented. Strunk et al. (2016) found positive effects in ELA achievement were concentrated in LAUSD schools that implemented more disruptive reform models. Zimmer, Henry, & Kho (2017) found significant increases in math, ELA, and science achievement were concentrated in Innovation Zones (iZones), a type of reform in which low performing schools remain in their local education agencies but semiautonomous districts-within-districts are created to help schools attract, retain and develop high quality teachers and leaders.

In contrast to the positive findings above, three studies found null and/or negative effects of school turnaround on student achievement (Dougherty & Weiner, 2017; Hamelt & Jacob, 2017; Henry & Harbatkin, 2018). In Michigan, waiver-based reforms in Priority Schools, defined as the state’s lowest performing 5 percent of schools, had little to no effect on math and ELA scores (Hamelt & Jacob, 2017). In Rhode Island, low performing schools that were required to implement more interventions experienced negative effects in ELA, while low performing schools that were required to implement fewer interventions experienced null effects in both ELA and math (Dougherty & Weiner, 2017). Lastly, Henry and Harbatkin (2018) found that the post-RttT state turnaround initiative in North Carolina had a negative impact on student test score growth across math, ELA, and science and across both year 1 and year 2 of the intervention.

Although the studies of school turnaround highlighted in Table 1 have increased understanding regarding which turnaround initiatives or models impact student achievement, they generally have examined the effects of school turnaround on student achievement in grade 3 and above. While the analyses of Strunk et al. (2016) included student achievement in ELA starting in grade 2, no studies as far as we are aware have analyzed the impact of school turnaround efforts specifically on early-grade student achievement. This lack of turnaround research on student achievement in early elementary grades is likely related to federal accountability requirements. Specifically, schools are not required to administer standardized assessments until grade 3 under federal law (Every Student Succeeds Act, 2015). As a result, the content (reading/ELA, math, or both), form (diagnostic or summative), timing, and data reporting requirements associated with standardized testing in grades K-2 varies widely by state (Croft, 2016). Our study aims to fill in this gap in the literature by estimating the effect of school turnaround on K-3 student achievement growth in North Carolina, as described further below.

North Carolina Transformation Initiative

The North Carolina Transformation (NCT) school turnaround initiative was implemented in 75 low-performing schools across the state during the 2015-16 and 2016-17 school years. 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 any of the prescribed turnaround models. The intervention was overseen by the North Carolina Department of Public Instruction (DPI) and their associated District and School Transformation (DST) team. Figure 1 graphically displays the "theory of change" for the NCT intervention.

Figure 1 ABOUT HERE

Services under NCT began with a Comprehensive Needs Assessment (CNA), which involved DST staff conducting interviews and observations in treatment schools to identify the strengths and weaknesses of the school and assess where supports should be targeted. CNA findings were then "unpacked" or discussed with treatment school staff. Unpacking discussions were usually held over multiple days and involved reviewing the CNA findings, conducting a "root cause analysis" that identified the causes underlying issues at the school, and conducting a "brown paper planning" activity which visually displayed the school improvement process. Following the CNA and unpacking, schools created their School Improvement Plans (SIP) which outlined their priorities and goals. Schools then submitted their SIPs through an online program called NCStar in order to receive feedback from DST coaches.

Based on the CNA, unpacking, and SIP, coaches were assigned to NCT schools with the goal of building school capacity. School transformation coaches (STCs) were assigned to work with principals, and instructional coaches (ICs) were assigned to work with teachers. Under NCT, there were no formal or state-mandated coaching requirements; instead, coaches were directed to assist in meeting individual school, principal, and teacher needs. Further, not all NCT schools were assigned both STCs and ICs, and there was large variation in the number and content of coaching visits by school. Based on the theory of change, the planning along with school transformation and instructional coaching was expected to lead to changes in principal and teacher practices, outcomes, and retention. In turn, student outcomes were expected to improve.

Henry and Harbatkin (2018) analyzed the effects of NCT on students test score growth in grades 3 and above using end-of-grade (EOGs) and end-of-course (EOCs) exams. They found the effect on students test score growth was -0.15 standard deviations in year one and -0.18 in year two of NCT. The authors assert that these negative effects are possibly due to the NCT initiative trying to serve all schools in the lowest performing 5 percent of schools, leading resources to be spread too thin. Given limited resources at DST, instructional coaches were not placed in every treatment school and the amount of coaching varied within schools. Thus, providing limited, inconsistent supports in these schools may have contributed to an already unstable school environment (Henry & Harbatkin, 2018). Based on these findings from grades 3 and above, it seems plausible that the effect of the NCT initiative on K-3 student reading growth measured using the DIBELS assessment will be negative or null but not likely positive.

Method

Data

This study relies on two sources of data. First, we utilize statewide administrative data from a longitudinal database 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. We specifically use data from the 2014-15, 2015-16, and 2016-17 school years for the analysis. We then merge the administrative data with mCLASS: Dynamic Indicators of Basic Early Literacy Skills Next (DIBELS) student achievement data. DIBELS, a universal assessment that measures the development of basic early literacy skills (Good et al., 2013), was administered statewide in North Carolina K-3 classrooms across the 2014-15, 2015-16, and 2016-17 school years. The DIBELS assessment was administered three times per school year, specifically at the beginning, middle, and end of the school year.

Analytic sample

The sample includes the 181 North Carolina schools that enrolled K-3 students and were eligible for treatment under NCT. Schools were excluded from NCT eligibility if they had a school performance grade (SPG) of C or above for the 2014-2015 school year, exceeded growth, were situated in one of the 10 largest school districts in the state or in Halifax County (which participated in a district-level turnaround during the same time as the NCT intervention), or were designated as a special or charter school. Of these 181 eligible schools, 39 were assigned to treatment based on a forcing variable (as further described in the section titled 'Assignment variable' below). However, noncompliance occurred on both sides of the treatment cutoff. In order for schools to receive services under NCT, district offices had to provide agreement. In a few instances, district officials requested that a school above the threshold receive services rather than or in addition to a school below the threshold. As a result, 33 of the 39 schools below the threshold received NCT services, six schools below the threshold declined services, and three schools above the threshold received services. In total, 36 schools that enrolled K-3 students received treatment under the NCT intervention.

Sample school characteristics are displayed in Table 2. Both treatment and comparison schools were often located in rural areas due to schools in the state's 10 largest districts being excluded from eligibility. Treatment schools had, on average, lower rates of fully licensed teachers (p<.001), higher rates of minority students (p<.001), and higher rates of economically disadvantaged students (p<.01) compared to comparison schools.

Table 2 ABOUT HERE

Outcome measures

We estimate the effects of NCT on student reading scores in grades K-3. We specifically operationalize reading scores as the end-of-year composite DIBELS score, which provides the best overall estimate of a student's reading proficiency (Good et al., 2013). Reading scores were standardized by grade and year.

Assignment variable

The state of North Carolina assigned schools to participate in the NCT intervention based on the 2014-15 school performance composite, a measure that represents EOY grade-level proficiency (GLP) in grades 3 and above. The cutoff score for NCT participation was 31.1 for schools enrolling K-3 students, with schools scoring below 31.1 being targeted for services. The performance composite is centered at this threshold.

Controls

Student reading scores from the beginning of the school year, school covariates, and student covariates were controlled for throughout the analysis to increase precision. School covariates include minority percentage, free or reduced lunch percentage, per-pupil expenditures (PPE) and PPE squared, and average daily membership (ADM) and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, limited English proficiency (LEP) (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

Empirical Strategy

We estimate the effect of being just below the threshold for assignment to NCT on K-3 student achievement using a fuzzy regression discontinuity (RD) design, which exploits the jump in probability of assignment at the cutoff (Imbens & Lemieux, 2007). We specifically estimate a two-stage least squares (2SLS) model, in which we instrument receipt of treatment by treatment eligibility, to account for noncompliance with treatment assignment. This approach allows us to estimate the average treatment effect on the treated at the margin of assignment to treatment, otherwise referred to as a local, compliance-adjusted treatment effect. To model the effect of NCT around the cutoff, we estimate a series of locally weighted linear regressions with a triangular kernel. In order to obtain first-stage z-statistics that are greater than four, which suggests that the forcing variable is a strong predictor of treatment under the guidelines for fuzzy

RD proposed by Deke et al. (2015), no bandwidths are specified. We also pool reading scores from all K-3 students across the 2015-16 and 2016-17 school years to maximize statistical power.

The first stage model is estimated as follows:

$$NCT_{is} = \alpha_0 + \alpha_1(GLP_s \leq 0) + \alpha_2 GLP_s + \alpha_3 y_{isBOY} + \gamma \boldsymbol{S}' + \pi \boldsymbol{C}' + \epsilon_{is} ,$$

where *NCT* represents participation in NCT for student *i* in school s, *GLP* is the centered forcing variable, $GLP \leq 0$ indicates assignment to NCT, y_{isBOY} is the reading score for student *i* in school *s* at the beginning of the school year, $\boldsymbol{S'}$ is a vector of school-level covariates, $\boldsymbol{C'}$ is a vector of student-level covariates, and ϵ is an idiosyncratic error term.

Further, we estimate the second stage model as follows:

$y_{isEOY} = \beta_0 + \beta_1 \widehat{NCT_{is}} + \beta_2 GLP_s + \beta_3[(GLP_s \leq 0) \times GLP_s] + \beta_4 y_{isBOY} + \gamma \boldsymbol{S}' + \pi \boldsymbol{C}' + \epsilon_{is}$, where *y* represents the reading score for student *i* in school *s* at the end of the year and $\widehat{NCT_{is}}$ is the predicted value of compliance with assignment to NCT. The interaction between the treatment eligibility variable and forcing variable allows for a different slope on either side of the cutoff. We estimate our main student achievement models with linear splines on either side of the cutoff but also include models with quadratic splines in the appendix. β_1 is the coefficient of interest, representing the estimated discontinuity at the cutoff. We also control for school and student covariates in the first and second stage models to increase precision. We pool all grade levels in our main specification models but also include separate models for kindergarten, first grade, second grade, and third grade in the appendix. Standard errors are clustered at the school level.

We also estimate the model 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 due to coaching not beginning until spring 2016. The 2017 estimate represents the effect of a full school year of coaching services. Because we include reading scores from the beginning of the school year on the right-hand side of both the first and second stage models, the outcome represents growth over one school year.[1]

Results

We find that the NCT intervention had no effect on student achievement growth in either 2016 or 2017. Figure 2 provides a graphical representation of these results with a linear specification. The vertical distance between the fit lines on either side of the cutoff shows the difference in outcomes associated with being in a school assigned to treatment. Across both the pooled and individual year model specifications, there is no visible effect of treatment in either 2016 or 2017.

Figure 2 ABOUT HERE

Table 3 displays the statistical results from the fuzzy RD models. Models 1-3 provide the pooled 2016 & 2017 estimates, Models 4-6 provide the individual 2016 estimates, and Models 6-8 provide the individual 2017 estimates. Models 1, 4, and 6 show estimates from fuzzy RD models using triangular kernels on the full sample with no bandwidths and accounting for the student's reading score from the beginning of the year in order to allow for a value-added interpretation of the effects. To increase precision, Models 2, 5, and 8 include school covariates, and Models 3, 6, and 9 include school and student covariates. As the graphical results suggest, the estimated coefficients associated with treatment are non-significant across all model specifications. The first-stage z-statistics are greater than four across all models, which suggests

[1] We also estimate models that utilize lagged student reading scores from the end of the previous school year and find similar results (see Table A-5 in the appendix).

that the forcing variable is a strong predictor of treatment under the guidelines for fuzzy RD (Deke et al., 2015).

Table 3 ABOUT HERE

As a robustness check, we estimate the effect of NCT on student reading score growth using a sharp RD design. We again find null effects across all model specifications in 2016 and 2017 (Table A-1), suggesting that the null effects are not being driven by systematic bias of those schools selected into or out of the NCT intervention. Our results are also robust to alternative specifications between the forcing and outcome variable. Specifically, we find null effects across both fuzzy and sharp RD models utilizing a quadratic spline (Table A-2 and A-3). However, the first stage of the fuzzy RD is not sufficiently strong across most quadratic spline model specifications. Under the guidelines for fuzzy RD (Deke et al., 2015), the first-stage test statistic should be a minimum of four to consider the instrument sufficiently strong. We denote models with weak first stages using a red box around the test statistic. The null effects of NCT also appear to hold across grade levels (Table A-4). We also find similar results when we estimate reading score growth using lagged reading scores from the end of the previous school year, which suggests that the null results are robust to alternative measures of lagged reading scores (Table A-5).

Validity Checks

In this section, we describe the six core assumptions of the RD design and then provide evidence that the data in this study meet those assumptions. The first assumption to the validity of the RD design is that there should be no manipulation of the forcing variable. Because the state of North Carolina determined the cutoff score on the assignment variable after schools administered end-of-year exams, manipulation of the forcing variable by schools is highly unlikely. Nevertheless, below we demonstrate both the graphical and statistical integrity of the forcing variable. Specifically, Figure 3 shows the density of the forcing variable across all eligible schools. The dashed vertical line at zero represents the cutoff score. The lack of a difference in density around the cutoff score demonstrates that there was no manipulation of the forcing variable. We also conducted a McCrary test to test the assumption of no manipulation. The test fails to reject the null of continuity of the density of the forcing variable (p=.4109), providing further evidence that the value of the school performance composite was not manipulated to influence treatment assignment near the cutoff.

Figure 3 ABOUT HERE

The second assumption to the validity of the RD design is that the functional form of the relationship between the outcome and forcing variable is correctly specified on both sides of the cutoff value. We estimate separate local linear regressions on either side of the cutoff to meet this condition. Figure 2 visually demonstrates that the relationship between the outcome and forcing variable is linear in both 2016 and 2017. We also estimated effects across both fuzzy and sharp RD models utilizing a quadratic polynomial fit and find that our results are robust to these alternative functional form specifications (Table A-2 and A-3). However, the first stage of the fuzzy RD is not sufficiently strong across most quadratic spline model specifications, as indicated by a z-test statistic below 4 (red box around the test statistic).

The third assumption of the fuzzy RD design assumes that eligibility for treatment is a strong predictor of participation in treatment. To meet this condition, Figure 4 shows the proportion treated by the forcing variable. Schools below the cutoff value of zero had a high probability of participation in the NCT intervention, whereas schools above the cutoff had a low probability of participation. First-stage z statistics from our preferred fuzzy RD models (see

Table 3) are also above the suggested minimum of four (Deke et al., 2015), further demonstrating that treatment eligibility is a strong predictor of compliance with treatment assignment.

Figure 4 ABOUT HERE

The fourth assumption of the fuzzy RD design estimated using 2SLS is that the treatment indicator meets the exclusion restriction. Biased estimates due to confounding variables are possible in our analysis because district offices had to provide agreement in order for schools to receive services under NCT. As a result, district officials requested that some schools above the threshold receive services and some schools below the threshold not receive services. If district officials requested certain schools receive or not receive services based on the perceived difficulty of increasing their performance (e.g. requested services in schools they perceived would be to difficult to improve and declined services in schools they thought would more easily improve), then our fuzzy estimates estimated using 2SLS would be biased downward. While we cannot explicitly test the exclusion restriction, we find null effects in both our preferred fuzzy RD models estimated using 2SLS (Table 3) and our sharp RD models that are not dependent on the exclusion restriction due to the intent to treat nature of this analysis (Table A-1). These similar results suggest that the 2SLS estimates are not biased due to failing the exclusion restriction.

The fifth assumption for the consistency of the fuzzy RD estimates is that the relationship between the forcing variable and outcome should be consistent in the absence of the intervention. This assumption cannot be tested directly because we cannot observe outcomes for treatment schools in the absence of treatment. Nevertheless, below we provide two indirect tests of the continuity of the outcome-forcing variable. First, we test the baseline equivalence of key covariates related to student reading scores across the treatment and comparison samples, conditional on the forcing variable. As shown in Table 4, the p-values associated with the key school-level student demographics, teacher demographics, and school performance covariates are all non-significant. Such non-significant results suggest that our treatment and comparison samples are balanced on observable characteristics and that the assumption of continuity of the outcome-forcing variable in the absence of treatment likely holds. Second, we examine the continuity assumption through creating a series of placebo cutoffs above and below the threshold and testing for discontinuities. None of the estimates associated with the placebo cutoffs are statistically significant in either 2016 or 2017 (Table A-6).

Table 4 ABOUT HERE

Lastly, the sixth assumption of the fuzzy RD is that there is no differential attrition across the treatment and comparison samples. Across the 2016 and 2017 years of this study, only one school in the comparison sample closed. As shown in Table 5, we estimated overall and differential levels of attrition at the school level using a sharp RD and controlling for the forcing variable. We find that the overall and differential levels of attrition are considered low based on the cautious boundary established by the What Works Clearinghouse (2017).

Table 5 ABOUT HERE

Discussion

This paper contributes to the literature on school turnaround through examining the effect of efforts to improve the lowest performing schools on K-3 student achievement. We find that the NCT intervention had null effects on K-3 reading score growth across both the 2016 and 2017 school years. These results are robust to fuzzy and sharp RD estimation strategies and alternative functional form specifications.

There are three plausible explanations for the lack of significant effects on K-3 student reading growth. First, it is possible that school leaders strategically focused NCT reform efforts on improving student performance in upper elementary grades due to being evaluated on student test scores in grades 3 and above, leading to null effects on early-grade student achievement. Such strategic practices by leaders of low-performing schools have been found in other studies of school accountability and turnaround (Cohen-Vogel, 2011; Grissom, Kalogrides, & Loeb, 2017). Next steps for this study include exploring the qualitative site visit data collected as part of the evaluation of NCT to examine whether reform efforts were targeted towards specific grades, including whether teachers in early grades received less instructional coaching than teachers in upper elementary grades. Second, the limited sample of schools that enroll K-3 students could possibly not allow for enough power to detect an effect that is small in magnitude. Despite including both school and student covariates in our models, most of our effect estimates are still imprecise. Third, it is possible that the intended effects of NCT may not have immediately translated into test score growth (see, e.g., Carlson & Lavertu, 2018). While we cannot know for certain whether the first two years of NCT laid the groundwork for improvement in future years, we find no evidence that delayed positive effects are emerging. For example, the NCT theory of change focused largely on building the capacity of individual teachers and principals, but Henry and Harbatkin (2018) found that many of those teachers left NCT schools in 2017, taking any increased capacity with them. Our null findings suggest that analyzing the impact of school turnaround efforts on student achievement in early elementary grades merits further study.

References

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

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

Cohen-Vogel, L. (2011). “Staffing to the test”: Are today’s school personnel practices evidence based?. *Educational Evaluation and Policy Analysis*, *33*(4), 483-505.

Croft, Michelle. (2016). *State Adoption and Implementation of K-2 Assessments.* Issue Brief, ACT Research & Policy. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/5738_Issue_Brief_State_Adoption_of_K-2_Assess_WEB_secure.pdf

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

Deke, J., Cook, T., Dragoset, L., Reardon, S., Titiunik, R., Todd, P., … Wadell, G. (2015). *Preview of Regression Discontinuity Design Standards.* What Works Clearinghouse.

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

ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2013). *DIBELS Next Technical Manual.* Eugene, OR: Dynamic Measurement Group, Inc.

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

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

Henry, G., & Harbatkin, E. (2018). The Next Generation of School Reform: Improving the Lowest Performing Schools without Disrupting the Status Quo. Presented at the 2018 Association for Education Finance and Policy (AEFP) conference.

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

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

Schueler, B. E., Goodman, J. S., & Deming, D. J. (2017). Can States Take Over and Turn Around School Districts? Evidence From Lawrence, Massachusetts. Educational

Evaluation and Policy Analysis, 39(2), 311–332. https://doi.org/10.3102/0162373716685824

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.* Retrieved from 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. Summary of School Turnaround Studies

Study	*Location*	*Grade levels included in analysis*	*Impact on student achievement*
Bonilla & Dee, 2017	Kentucky	3-8	+ Positive effect, gap group students, ELA & math
Carlson & Lavertu, 2018	Ohio	3-8	+ Positive effect, ELA
Dougherty & Weiner, 2017	Rhode Island	3-8	- Negative effect, Focus schools, ELA; Null effect, Warning schools, ELA & math
Hemelt & Jacob, 2017	Michigan (Priority Schools)	3-8	Null effect, ELA & math
Henry & Harbatkin, 2018	North Carolina	3-12	- Negative effect, ELA, math, & science;
Papay & Hannon, 2018	Massachusetts	3-8	+ Positive effect, ELA & math
Schueler, Goodman, & Deming, 2017	Lawrence, Massachusetts	3-8, 10	+ Positive effect, ELA & math
Strunk et al., 2016	Los Angeles, California	2-11	Null effect, first cohort; + positive effect, second cohort, ELA; - negative effect, third cohort, ELA & math
Sun, Penner, & Loeb, 2017	San Francisco, California	3-12	+ Positive effect, ELA & math
Zimmer, Henry, & Kho, 2017	Tennessee	3-12	+ Positive effect, iZone, ELA, math, & science; Null effect, Achievement School District

Table 2. School sample characteristics

	Treatment	Comparison	Sig.
Urbanicity			
City	0.0 (0.00)	0.0 (0.20)	
Suburb	0.0 (0.00)	0.1 (0.23)	
Town	0.0 (0.17)	0.1 (0.31)	
Rural	1.0 (0.17)	0.8 (0.40)	*
Student achievement			
2015 performance composite (centered)	-4.0 (4.24)	10.4 (5.78)	***
Teacher qualifications			
Novice teacher percent	30.8 (16.00)	27.4 (12.86)	
Fully licensed teacher percent	92.9 (9.07)	96.5 (4.12)	***
Student demographics			
Minority percent	88.4 (9.66)	63.9 (22.08)	***
Economically disadvantaged percent	86.4 (11.03)	79.2 (13.93)	**
Per pupil spending	9696.8 (1589.43)	9988.4 (1523.54)	
Average daily membership	444.2 (160.11)	409.0 (168.46)	
N	36	145	

Means and standard deviations are presented

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

Table 3. Fuzzy RD results *(outcome=reading score growth)*

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	0.031	0.037	0.056	0.054	0.075	0.077	0.007	-0.004	0.019
	(0.0759)	(0.0664)	(0.0569)	(0.0936)	(0.0868)	(0.0693)	(0.0726)	(0.0641)	(0.0628)
N	86528	86528	86528	43987	43987	43987	42541	42541	42541
First-stage Z	6.09	6.19	6.17	5.74	5.85	5.84	6.42	6.60	6.58
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Robust standard errors clustered at the school level. Estimates from fuzzy RD using triangular kernel on full sample, no bandwidths, and linear spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table 4. Sample balance conditional on forcing variable

	Treatment	Comparison	p-value
School-level student demographics			
ED percent	88.99	87.98	0.807
Minority percent	87.73	74.14	0.088
Black percent	67.41	48.37	0.237
Hispanic percent	13.90	24.71	0.242
ADM	411.93	441.66	0.787
Teacher demographics			
Novice teacher rate	0.38	0.45	0.223
Fully licensed teacher rate	0.94	0.95	0.850
School performance			
School EVAAS	-3.69	-2.11	0.638

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

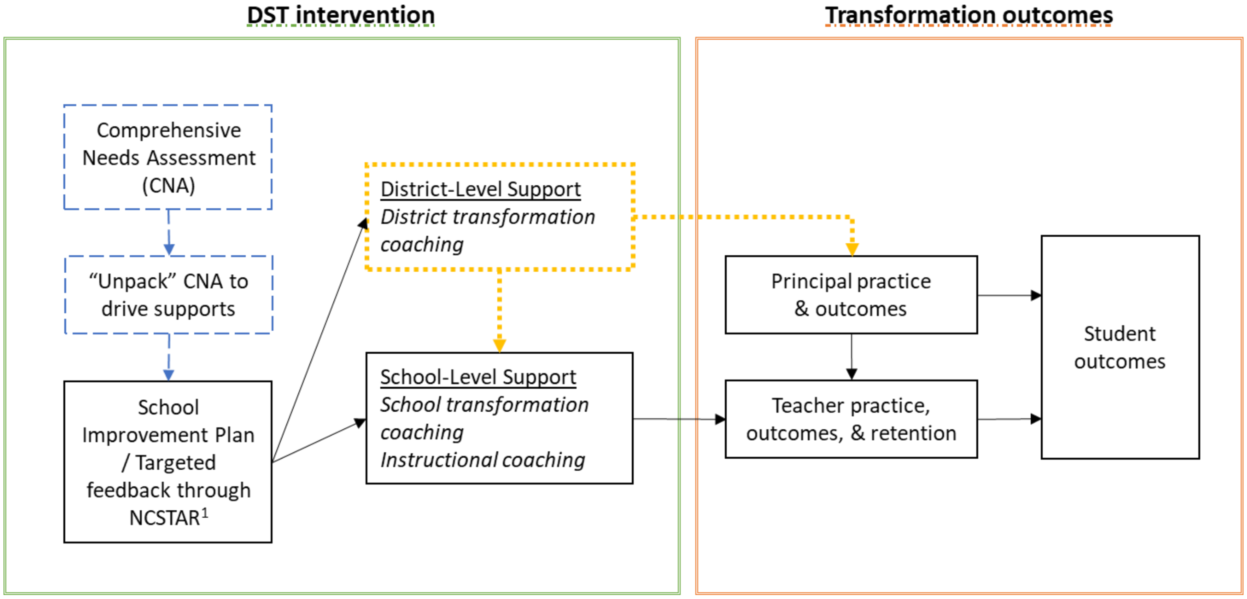
Table 5. Attrition

	2016 & 2017
β_{treat}	0.043
$\beta_{compare}$	0.000
$\beta_{overall}$	0.0215
β_{diff} (SE)	-0.043 (0.0276)

Estimates from RD predicting attrition at the school level and controlling for the forcing variable

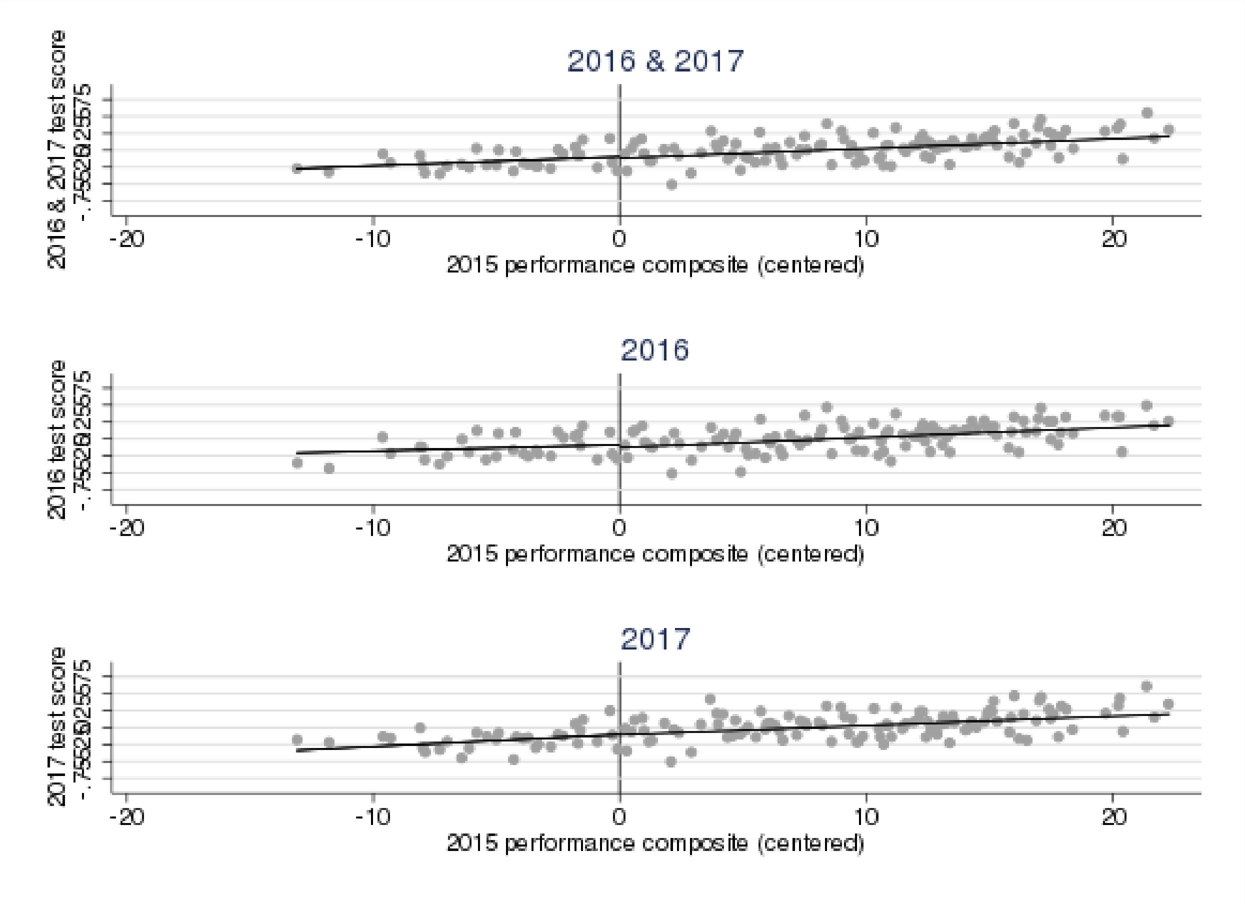
Figures

Figure 1. North Carolina Transformation Theory of Change


DST intervention
Transformation outcomes
Comprehensive Needs Assessment (CNA)
"Unpack" CNA to drive supports
School Improvement Plan / Targeted feedback through NCSTAR[1]
District-Level Support
District transformation coaching
School-Level Support
School transformation coaching
Instructional coaching
Principal practice & outcomes
Teacher practice, outcomes, & retention
Student outcomes

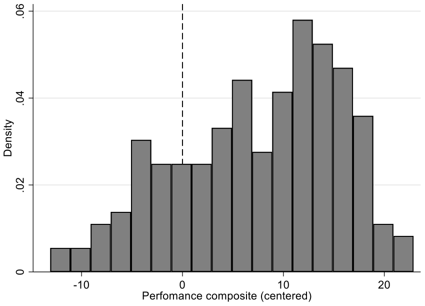
NOTE: Blue dashed lines indicate activities in which the timeline varies by treatment schools. Yellow dotted lines indicate activities not available to all districts.

Figure 2. Reading score growth by distance from assignment threshold



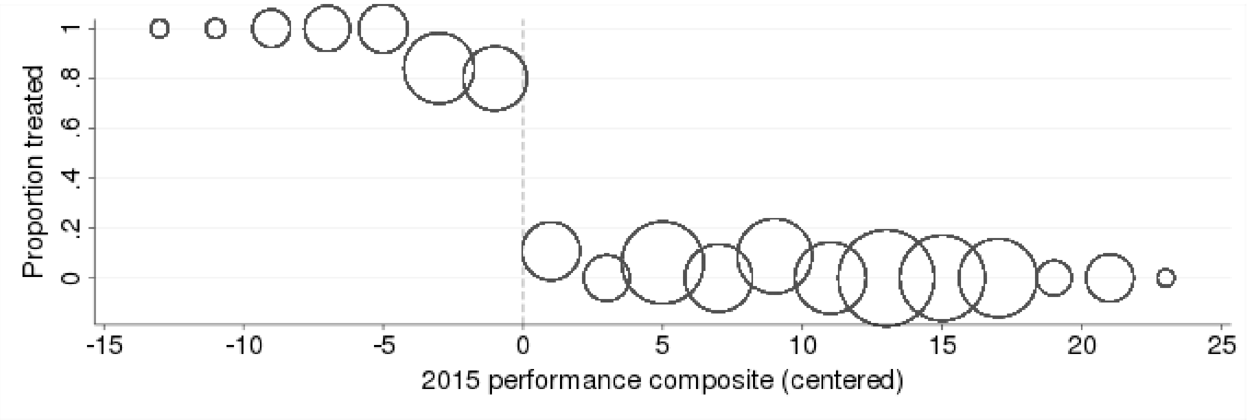
NOTE: Markers represent bin averages and line is linear fit. Estimation using triangular kernel, with average bin width of .011 to left of cutoff and .031 to right of cutoff in 2016 and 2017, .046 to left of cutoff and .050 to right of cutoff in 2016, and .025 to left of cutoff and .064 to right of cutoff in 2017.

Figure 3. Graphical integrity of the forcing variable



NOTE: Bin width is 2. Includes all eligible schools

Figure 4. Proportion treated by forcing variable



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

Appendix

Table A-1. Sharp RD results *(outcome=reading score growth)*

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ITT	0.022	0.025	0.037	0.037	0.050	0.050	0.005	-0.003	0.013
	(0.0524)	(0.0443)	(0.0366)	(0.0629)	(0.0559)	(0.0431)	(0.0519)	(0.0444)	(0.0427)
N	86528	86528	86528	43987	43987	43987	42541	42541	42541
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Robust standard errors clustered at the school level. Estimates from sharp RD using triangular kernel on full sample, no bandwidths, and linear spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table A-2. Fuzzy RD results with quadratic spline *(outcome=reading score growth)*

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	-0.014	-0.005	0.031	-0.061	-0.050	-0.022	0.034	0.039	0.080
	(0.1076)	(0.0947)	(0.0767)	(0.1278)	(0.1205)	(0.0899)	(0.1106)	(0.0996)	(0.0991)
N	86528	86528	86528	43987	43987	43987	42541	42541	42541
First-stage Z	3.88	3.98	3.95	3.54	3.65	3.58	4.26	4.45	4.44
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Red box denotes first-stage z statistics that are less than the Deke et al. (2015) suggested threshold value of 4 for a sufficiently strong first stage. Robust standard errors clustered at the school level. Estimates from fuzzy RD using triangular kernel on full sample, no bandwidths, and quadratic spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table A-3. Sharp RD results with quadratic spline ***(outcome=reading score growth)***

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ITT	-0.009	-0.003	0.019	-0.038	-0.030	-0.013	0.023	0.025	0.051
	(0.0707)	(0.0596)	(0.0458)	(0.0824)	(0.0745)	(0.0526)	(0.0731)	(0.0630)	(0.0603)
N	86528	86528	86528	43987	43987	43987	42541	42541	42541
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Robust standard errors clustered at the school level. Estimates from sharp RD using triangular kernel on full sample, no bandwidths, and quadratic spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table A-4. Fuzzy RD results by grade level ***(outcome=reading score growth)***

Panel A: Kindergarten

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	0.089	0.079	0.082	0.129	0.137	0.102	0.048	0.019	0.021
	(0.1765)	(0.1556)	(0.1383)	(0.2009)	(0.1816)	(0.1651)	(0.1831)	(0.1668)	(0.1522)
N	20075	20075	20075	10229	10229	10229	9846	9846	9846
First-stage Z	6.25	6.33	6.31	5.69	5.81	5.92	6.76	6.87	6.88
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Panel B: First Grade

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	0.041	0.048	0.054	0.085	0.114	0.099	-0.005	-0.020	0.005
	(0.0978)	(0.0961)	(0.0808)	(0.1436)	(0.1413)	(0.1140)	(0.0872)	(0.0910)	(0.0854)

N	21435	21435	21435	11040	11040	11040	10395	10395	10395
First-stage Z	6.25	6.33	6.31	5.69	5.81	5.92	6.76	6.87	6.88
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Panel C: Second Grade

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	-0.041	-0.026	0.008	-0.061	-0.037	-0.010	-0.024	-0.025	0.013
	(0.0694)	(0.0706)	(0.0615)	(0.0739)	(0.0764)	(0.0690)	(0.0843)	(0.0830)	(0.0747)
N	21756	21756	21756	11011	11011	11011	10745	10745	10745
First-stage Z	5.92	6.03	6.04	5.42	5.56	5.56	6.42	6.60	6.60
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Panel D: Third Grade

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	0.008	0.031	0.053	0.038	0.074	0.090	-0.023	-0.015	-0.004
	(0.0709)	(0.0686)	(0.0683)	(0.0967)	(0.0963)	(0.0843)	(0.0897)	(0.0899)	(0.0971)
N	23262	23262	23262	11707	11707	11707	11555	11555	11555
First-stage Z	5.95	6.07	6.02	6.11	6.20	6.13	5.75	5.95	5.95
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Robust standard errors clustered at the school level. Estimates from sharp RD using triangular kernel on full sample, no bandwidths, and linear spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table A-5. Fuzzy RD results with lagged reading score from the end of the previous school year *(outcome=reading score growth)*

	2016 & 2017			2016			2017		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOT	0.020	0.019	0.031	0.061	0.073	0.065	-0.026	-0.042	-0.028
	(0.0565)	(0.0520)	(0.0499)	(0.0754)	(0.0672)	(0.0587)	(0.0656)	(0.0678)	(0.0679)
N	59570	59570	59570	30485	30485	30485	29085	29085	29085
First-stage Z	6.11	6.22	6.19	5.83	5.94	5.91	6.38	6.55	6.52
School covariates		X	X		X	X		X	X
Student covariates			X			X			X

Robust standard errors clustered at the school level. Estimates from sharp RD using triangular kernel on full sample, no bandwidths, and linear spline. All models include lagged student reading scores (from the end of the previous school year) on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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

Table A-6. Placebo estimates from fuzzy RD *(outcome=reading score growth)*

Panel A: 2016 & 2017

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Placebo Cutoff	-8	-6	-4	-2	2	4	6	8
TOT	-3.208	0.158	-0.015	-0.056	0.057	-0.083	-0.092	0.014
	(6.0691)	(0.2494)	(0.1636)	(0.1085)	(0.1234)	(0.7904)	(0.2807)	(0.1393)
Observations	86528	86528	86528	86528	86528	86528	86528	86528

Panel B: 2016

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Placebo Cutoff	-8	-6	-4	-2	2	4	6	8
TOT	-2.323	0.304	0.214	0.030	0.228	-0.281	0.176	0.138
	(2.6138)	(0.3788)	(0.2018)	(0.1198)	(0.2028)	(0.8349)	(0.3802)	(0.1820)
Observations	43987	43987	43987	43987	43987	43987	43987	43987

Panel C: 2017

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Placebo Cutoff	-8	-6	-4	-2	2	4	6	8
TOT	-11.141	-0.060	-0.240	-0.278	-0.117	0.773	-0.391	-0.139
	(84.8533)	(0.2707)	(0.2348)	(0.2070)	(0.1415)	(3.7136)	(0.3658)	(0.1475)
Observations	42541	42541	42541	42541	42541	42541	42541	42541

Robust standard errors clustered at the school level. Estimates from fuzzy RD using triangular kernel on full sample, no bandwidths, and linear spline. All models include student reading scores from the beginning of the school year on the right hand side. School covariates include minority percentage, free or reduced lunch percentage, PPE and PPE squared, and ADM and ADM squared. Student covariates include grade level with kindergarten as the reference category, gender, ethnicity with white as the reference category, student with disabilities (current), academically gifted, LEP (current), overage, assessed by classroom teacher at beginning of school year, assessed by classroom teacher at end of school year, student mobility, and days between beginning and end of year assessments.

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