Student Matching and Tracking under Accountability: Evidence from No Child Left Behind

Michael Gilraine*

February 15, 2016

VERY PRELIMINARY AND INCOMPLETE.

PLEASE DO NOT CITE.

ABSTRACT

One of the key aims of policymakers is to close the achievement gap between high- and low-performing children, especially among minority and non-minority children. To accomplish this goal, the No Child Left Behind Act of 2001 (NCLB) set subgroup-specific targets that schools had to meet in order to avoid sanctions. Using a rule embedded in NCLB that subgroup-specific targets only count towards a school's accountability performance if there are forty or more students in that subgroup, I shed light on the efficacy of this policy by using a regression discontinuity design that compares schools just below relative to just above forty students in a given subgroup. The results show that students in subgroups facing accountability pressure receive a significant boost to their test scores: about a 0.10σ and 0.02σ improvement in math and English test scores, respectively. Further analysis shows that school administrators attain these results by tracking these students into classes with higher-achieving peers and matching these students to higher value-added teachers. These results shed light on how accountability pressure can change the tracking and matching systems that schools use, providing policymakers with an additional lever to wield when employing accountability schemes that aim to reduce achievement gaps. *JEL* codes: I21, I24, I28.

Keywords: Matching; Tracking; Teacher Effects; Peer Effects; Regression Discontinuity.

*I would like to thank Robert McMillan for his guidance throughout this project. Also thanks to Philip Oreopoulos, Aloysius Siow, Adam Lavecchia, Uros Petronijevic, Mathieu Marcoux, Jessica Burley, and members of the University of Toronto's CEPA Seminar and Public Economics Workshop for their help, comments and advice. All remaining errors are my own. Contact: Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, Canada, M5S 3G7. Please send comments to mike.gilraine@mail.utoronto.ca.

1 Introduction

Policy interventions in education often aim not just to improve student achievement generally, but also to close the achievement gaps between students of different racial and socioeconomic backgrounds. To accomplish these goals, policymakers often explicitly include the narrowing of achievement gaps as a target for schools to reach. For example, one the three goals of the recent Race To The Top initiative is to "decrease the achievement gaps across student subgroups" (US Department of Education 2009, p.4).

I shed light on the efficacy of these policies by looking at the effect of the subgroup-specific accountability standards embedded in the No Child Left Behind Act of 2001 (NCLB). Under NCLB, schools were held accountable to not just overall levels of student achievement, but also to the level of student achievement in nine student subgroups. If schools did not meet one of the subgroup requirements, they were subject to school level sanctions.[1]

Using a regression discontinuity design that exploits a rule that determines whether a student subgroup is accountable under NCLB, I show that these subgroup-specific accountability measures yield large improvements for students in that subgroup. Specifically, I use the fact that a student subgroup is accountable under NCLB only if there are forty or more students in a given subgroup. The intuitive idea is that schools near a specific subgroup threshold have similar observable and unobservable characteristics, while their subgroup accountability pressure differs, providing the basis for utilizing a RD approach to identify causal impacts. I show that facing subgroup-specific accountability pressure increases the test scores of students within that subgroup significantly: by about a 0.10σ and 0.02σ improvement in math and English test scores, respectively.

Due to limited resources, however, schools may lower achievement gaps by appropriating resources earmarked to one student subgroup to another. In light of recent research

[1]These sanctions became progressively tougher if schools failed NCLB repeatedly. The sanctions ranged from developing a school improvement plan to implementing a school restructuring plan, often causing many staff to be replaced (Ahn and Vigdor, 2014).

highlighting the importance of teachers[2] and peers[3] to a student's achievement, schools may alter how students are assigned classes, by tracking them with higher achieving peers and matching them to better teachers.[4]

I show that schools held accountable for a subgroup improve test scores in that subgroup by changing their classroom assignments. Specifically, students are assigned classes with higher achieving peers and higher value-added teachers. This shows that schools respond to subgroup-specific accountability not by increasing subgroup-specific effort, but by shifting their resources of high-achieving peers and skillful teachers to these students.

These results help to shed new light on the manner in which students are tracked and matched.[5] In general, evidence points to parental pressure being an important determinant, with more involved parents being able to pressure school administrators to place their students in classes with better peers and teachers.[6] Other evidence suggests that teachers may pressure administrators to teach high-ability children (Finley, 1984). The end result of this assignment mechanism is clear: school administrators systematically match students of like ability together and systematically assign high-achieving students to better teachers (Kalogrides and Loeb, 2013).

The ability of accountability schemes like NCLB to alter classroom assignments gives policymakers another valuable tool that can be used to reduce the achievement gaps among students. For example, Fryer and Levitt (2004) suggest that the black-white test gap, which only appears after the first two years of school, is caused by differences in the schooling inputs that students face. While much of these differences in schooling inputs are caused by across school differences driven by decentralized parental school choice, a substantial amount of the

[2]See Rockoff (2004); Kane et al. (2013); Chetty et al. (2014b).

[3]See Sacerdote (2001); Zimmerman (2003); Carrell et al. (2009).

[4]I take these terms from Dieterle et al. (2015) who define tracking as students being systematically assigned with certain types of peers and matching where students are systematically assigned certain teachers.

[5]A related literature has looked at whether student tracking invalidates teacher value-added measures, a possible consequence of student tracking that is beyond the scope of this paper (Rothstein, 2010; Koedel and Betts, 2011; Chetty et al., 2014a).

[6]For example, Lareau (2000) discusses how middle class parents are knowledgeable about the abilities of teachers in their school and attempt to have their students placed in the class of their preferred teacher.

variation in schooling inputs is within schools, which are driven by centralized school administrators choices which may be more easily influenced by accountability schemes.[7] Modeling the classroom assignment mechanism and using the regression discontinuity estimates produced in this paper provide a path forward on structurally estimating the effects of various accountability schemes on the within-school distribution of resources among students, presenting a possible avenue for policymakes to narrow the persistent achievement gaps among students.[8]

The rest of the paper is organized as follows: The next section describes the NCLB subgroup rule. Section 3 provides an empirical framework for the study and discusses the data set used. Results are presented in Section 4 and are placed in a broader context in Section 5. Section 6 concludes.

2 Background

The No Child Left Behind Act of 2001 (NCLB), the major education initiative of the Bush administration, aimed to raised educational achievement. In pursuit of that goal, NCLB stated that if a school's proficiency rate was below some pre-set target the school would not meet Adequate Yearly Progress (AYP).[9] Schools not making AYP would be subject to progressively harsher corrective action, including school takeovers.[10] In North Carolina, proficiency designations were based on standardized tests administered at the end of a school year: if a student attained a pre-defined score on the end of grade test, the student would be labelled proficient. Test results for grades 3-8 were used for NCLB purposes in elementary and middle schools.[11]

One of the key purposes of NCLB was to "clos[e] the achievement gap between high- and

[7] See Nechyba (2006) for a more in-depth discussion of this point.

[8] Moving forward, this will become one of the major focuses of this paper.

[9] There were some exceptions written into the law that a school could meet AYP even if its proficiency rate was below the target. The two main exemptions were the confidence-interval and safe harbour exemptions.

[10] For a description of these sanctions and their effects, see Ahn and Vigdor (2014).

[11] High schools had additional targets, including graduation rates, to achieve.

lowperforming children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers" (No Child Left Behind Act of 2001, 115 STAT. 1440). The mechanism NCLB used to close these gaps are the subgroup-specific proficiency rates that schools must meet to pass AYP. Specifically, schools were required to meet both an overall school proficiency rate and subgroup-specific proficiency rates. The subgroup-specific targets were set at the same level as the overall target, which meant that most schools failing AYP did so through failing a subgroup-specific target rather than the overall target. The student subgroups are: black, hispanic, white, asian, multi-racial, indian, economically disadvantaged, limited English proficient and students with disabilities.

Since some schools have few students in a subgroup, making the ability to pass AYP hinge on a small subset of their student population, legislators incorporated a rule in the law that there must be forty or more students in a given subgroup for the school to be held accountable for that subgroup's proficiency target. Therefore any school with forty or more students in a subgroup faced accountability pressure for that subgroup, while schools with less than forty students did not. Crucially, the method in which NCLB counted students for this purpose was somewhat complicated by the fact that for a student's proficiency status to be included towards the calculation of the school's NCLB proficiency rate (overall or in a subgroup) that student must be present at the school for a minimum of 140 days. For this reason, schools often could have more than forty students in subgroup take the end of grade test but, since numerous students in that subgroup were at the school for less than 140 days, the school was not held accountable for that subgroup.

In addition, it is important to note that NCLB is a proficiency based scheme. For example, in 2003 a school needed to have 74.6 percent of its students overall, and in each subgroup with forty or more students, be labelled proficient to pass AYP.[12] Therefore, schools are not

[12]There are additional targets that the school also must attain to pass AYP, including having more than 95 percent of students taking the standardized math and English tests and have increasing attendance. In North Carolina, less than two percent of schools in my sample failed one of these requirements.

rewarded for student test score gains unless that test score gain brings the student over the proficiency threshold. This led to an incentive for schools to 'teach to the distribution' by focusing their efforts towards students near the proficiency threshold (Neal and Schanzenbach, 2010). Therefore, the subgroup-specific accountability provisions in NCLB may influence higher test score gains for students in a given subgroup who have predicted test scores near the proficiency threshold.

Finally, NCLB's proficiency based system has implications on dynamic gaming. In Macartney (2016), the author shows that schools under a value-added accountability scheme redirect their effort towards students in higher grades. This occurs because investments in lower grades increase the school's value-added targets the subsequent year since the student remains at the school. A test score improvement in a later grade, however, will not contribute to the school's value-added targets next year since the student will have moved on to a middle school. In a proficiency-based scheme, however, the dynamic gaming works in the opposite direction. To visualize this, consider the most common North Carolina school that has a K-5 grade span configuration. On one hand, given that test scores for grade 3-5 students count for NCLB, an investment in a grade K-3 student increases the probability that the school attains its proficiency target for three years, since the student will score better in the next three tests taken (grade 3, grade 4 and grade 5). Improving the test score of a grade 5 student, on the other hand, only increases the probability of the school reaching its proficiency target for one year, as all the tests the student takes after grade 5 are in a different school. Under this type of dynamic gaming, we would expect to see larger test score gains for K-3 students in a given subgroup rather than grade 5 students under NCLB accountability.

3 Empirical Framework

The effect of subgroup-specific accountability on student achievement is ultimately an empirical question. To answer it, I use a regression discontinuity design that takes advantage of the fact that only schools with more than 40 students in a given subgroup are required to meet the subgroup-specific accountability targets. The essence of the empirical strategy is to compare students' outcomes in a given subgroup in schools with slightly less than forty students to those with slightly more. To illustrate the idea, consider a school with thirty-nine disadvantaged students to one with forty disadvantaged students. It is unlikely that schools with thirty-nine disadvantaged students differ much from schools with forty disadvantaged students. However, due to the policy rule, the school with the forty disadvantaged students is held accountable and must have that disadvantaged subgroup reach the proficiency target, while the school with thirty-nine disadvantaged students does not.

As the schools are likely to be similar in other observable and unobservable dimensions, we can compare outcomes of disadvantaged students between the two schools to examine the effect of NCLB's subgroup-specific accountability provisions. As the cutoff expands, possible relationships between the number of students in a subgroup and student achievement may appear, making it necessary to control for the number of students in a subgroup within the school through some function. To keep with the spirit of comparing schools close to the cutoff, I restrict regressions to only include schools within a bandwidth of b_{lst} students of the forty student cutoff.

Specifically, I run the following regression for all students in subgroup l of school s at time t:

$$\begin{aligned} y_{ilst} = \alpha_0 + \tau_{SRD}\mathbb{1}\{l \geq 40\}_{lst} + \beta_1(\#\ in\ l)_{lst} + \beta_2(\#\ in\ l * \mathbb{1}\{l \geq 40\})_{lst} \\ + \delta_g + \lambda_t + \kappa_d + \phi X_{ilst} + \epsilon_{ilst}, \quad \forall lst \in [40 - b_{lst}, 40 + b_{lst}], \end{aligned} \tag{3.1}$$

where y_{ilst} represents the standardized test score of student i in subgroup l in school s at

time t, $\mathbb{1}\{l < 40\}_{lst}$ is an indicator variable equal to one if school s has more than forty students in subgroup l at time t (and thus is subject to accountability pressure for that subgroup), $Post_t$ is an indicator equal to one if NCLB has been implemented, $(\#\ in\ l)_{lst}$ is the number of students in subgroup l in school s at time t (which is interacted with the treatment indicator to allow for different functional forms on either side of the cutoff), X_{ilst} is a set of individual level controls and δ_g, λ_t and κ_d are grade, year and district fixed-effects, respectively.

Our coefficient of interest is the regression discontinuity estimate τ_{SRD}, which, under assumptions that will be tested in Section 4, represents the causal effect of NCLB subgroup-specific accountability pressure. All regression discontinuity estimates should be construed as local average treatment effects. Further, since schools adhere to the student count rule perfectly (see Section 4), this is a sharp regression discontinuity design.

3.1 Data

The data used for this study are administrative data from the North Carolina Education Research Center (NCERDC). This includes information on North Carolina students and teachers for the 2002-03 to 2004-05 school years. Given that NCLB was enacted for the 2002-03 school year, the data cover the first three years of NCLB.[13] The data set contains test scores for each student in mathematics and English for grades two through eight from a standardized test that is administered at the end of each school year in North Carolina.[14] Test scores are reported on a developmental scale. To create comparability of test scores across grades, I standardize this scale to have a mean of zero and a variance of one for each grade year.

The data set also has unique student and teacher identifiers, which allows students and teachers to be linked and tracked over time. Classroom assignment data is inferred by the

[13]Future versions of this paper will likely use data from additional years.

[14]With the notable exception of grade 2 tests, which are administered at the start of the grade three school year.

teacher who administers the test at the end of each school year. In elementary schools, this is almost always the student's classroom teacher.[15] In middle school, however, the teacher administering the test may not be the classroom teacher of the student for that subject, implying that reliable classroom assignments are only available until grade five.

Data on school level subgroup counts and what targets the school must meet to pass Adequate Yearly Progress are obtained through the AYP reports.[16] The AYP reports include detailed data on which subgroup-specific targets the school must meet to pass AYP and whether or not the target was achieved. The reports also record the number of students in each subgroup that are used for NCLB purposes, which, because of the 140 rule described in Section 2, often differ from the number of students in a given subgroup that are tested in a given school.[17] For this reason, the number of students in a given subgroup in a school is always calculated using the data from the AYP reports.

Summary statistics are reported in Table 1. Column (1) shows the student characteristics for all students in the sample. North Carolina has a majority white student population with a substantial black minority population. Hispanic and asian students make up less than ten percent of the student body. However, when we restrict to observations near the forty student threshold in column (2), we see substantial differences among these students relative to the student body as a whole. First, these students are much less likely to be white and far more likely to be minority, with the percentage of hispanic and asian students more than doubling. This is likely an artifact of the sample: schools are far more likely to have a minority subgroup close to the forty student subgroup threshold than a majority subgroup. Students in the RD sample are correspondingly more likely to be English learners than the whole sample.

Finally, summary statistics are also provided for the RD sample for grade 2 and 3 students

[15]This is the method of obtaining classroom assignment using NCERDC data that is prevalent in the literature (for example, see Clotfelter et al. (2006)).

[16]These reports are available at `http://accrpt.ncpublicschools.org/app/2003/ayp/`.

[17]These differences can be quite large: in some cases, schools had over fifty students in a given subgroup take the end of grade test, but the subgroup count was still below forty for NCLB purposes.

only. In Section 4, this sample is used to look at changes in schooling inputs as non-elementary grades do not have reliable classroom assignment information. These students generally perform worse on tests than their whole sample counterparts.[18] In general, they hold similar racial characteristics to the all grades RD sample, though have a substantially higher rate of free or reduced price lunch students, likely an artifact of the grade restriction as students in earlier grades are designated free or reduced price lunch at higher rates than their older counterparts.

4 Results

The validity of the regression discontinuity design introduced in Section 3 is discussed. Results are then presented, with these results being discussed in Section 5.

4.1 Tests of the Validity of the Regression Discontinuity Design

Because schools do not face any subgroup-specific accountability if they have below forty students in a subgroup, there is an incentive for schools to manipulate the number of students in a given subgroup in order to avoid being held accountable for the performance of that subgroup. As schools that manipulate their subgroup numbers to avoid subgroup-specific accountability may systematically differ from those that do not, such manipulation may invalidate the RD design.

Section 2 describes how the number of students in a given subgroup is calculated. Given the calculation procedure, schools can manipulate the number of students in a subgroup by either changing a student's subgroup designation, refusing admission to a student of a certain subgroup, or by preventing a student from reaching 140 days at the school through suspension or expulsion. For racial subgroup categories, changing a student's subgroup designation seems difficult, though there is some evidence that schools are able to change a

[18] This can be seen since their grade-year level standardized test scores are below zero.

student's disability designation. In light of that, the disability subgroup is omitted from the analysis that follows.[19]

To check whether schools manipulate the number of students in a subgroup through admission refusal or taking advantage of the 140 day rule, Figure 1 plots the distribution of school subgroups by the number of students in that subgroup. If schools were manipulating the number of students in a subgroup to avoid accountability pressure, we would expect there to be a large number of schools just to the left of the forty student threshold relative to the right. Visually, there does not appear to be any excess density around the threshold. A formal test of continuity in the density around the threshold (McCrary, 2008) confirms our visual analysis: the null hypothesis of continuity at the cutoff is not rejected.

Another method to determine the validity of the regression discontinuity design is to check for discontinuities in observable characteristics at the forty student cutoff. While observable characteristics can be controlled for, discontinuities in observable characteristics may be suggestive of changes in unobservables around the cutoff. Figure 2 graphs the mean covariates by the number of students in a subgroup. These covariates include race, free lunch status, limited English proficient status, gifted status, disability status, and parental education. For all covariates, smooth distributions around the cutoff are observed.

4.2 First-Stage

For the regression discontinuity design to be valid, crossing the forty student in a subgroup threshold must cause schools to be held accountable under NCLB. Figure 3 plots whether or not the subgroup-specific accountability target was used for NCLB purposes by the number of students in that subgroup. As expected, once the forty student threshold is crossed, schools are held accountable for the proficiency rate in that subgroup. In fact, adherence to the rule is perfect, denoting that there is a one-hundred percent jump in the probability that schools are held accountable for that subgroup once there are forty or more students in

[19]There is no evidence of manipulation of the other two non-racial categories of limited English proficient or free or reduced price lunch status.

that subgroup.[20]

4.3 Reduced-Form

Figure 4 indicates a reduced-form relationship between the number of students in a given subgroup and mean standardized test scores for math and English. For math, there is significantly higher achievement for students in a subgroup that is above the forty student threshold. Point estimates suggests that there is a 0.10σ improvement in math scores when schools are held accountable for the subgroup's performance, though this effect is not statistically significant without student controls. The effect sizes for English are much smaller in magnitude and are generally indistinguishable from zero.

Panel A of Table 2 reports the results from Figure 4, with student level controls being added in columns (2) and (4). With the addition of controls the estimated math effect of crossing the achievement threshold declines to 0.056σ but becomes statistically significant at the ten percent level. The effects on English remain statistically indistinguishable from zero.

In addition, I report results for grades 2 and 3 only. There are two reasons for this: first, classroom and teacher assignments are not available for the following year for students in grade 5 or later, implying that the analysis that follows is only able to use grades 2, 3, and 4. Second, research into the dynamics of educational accountability shows that schools time their inputs across grades to maximize the likelihood of passing the accountability standard (Macartney, 2016). As described in Section 2, in the case of a proficiency system like NCLB, this implies that administrators will invest in earlier grade students as these students will contribute to passing the targets in future years so long as they remain at the school.[21]

[20] In other words, this is a sharp regression discontinuity design. This perfect jump is not observed for limited English status students. This is likely due to the fact that North Carolina does not require newly designated limited English proficient students to be tested, implying that the NCLB student counts are larger than the counts used for the rule. For this reason, the limited English proficient subgroup is omitted from the results in this paper.

[21] This type of dynamic gaming is in the opposite direction of Macartney (2016), who investigated a value-added based accountability system. I am currently exploring the dynamics of NCLB accountability in

Figure 5 shows the relationship between test scores and the number of students in a subgroup for grades 2 and 3 only. As one can see, there is bigger jump in outcomes around the discontinuity, suggesting that the school may respond to the dynamic incentives of NCLB by investing more in earlier grades.[22] Finally, Columns (2) and (4) of Table 2, Panel B, report these estimates with covariates. We see that crossing the forty students threshold causes a 0.104σ improvement in test scores in math for grade 2 and 3 students, which is statistically significant at the five percent level.

4.4 School Inputs

Next, we look at possible reasons for the improvement in test scores by looking at the schooling inputs that students face in elementary school grades. Repeating the same regression discontinuity design described in Section 3, we replace the test score outcomes with measures of teacher quality and classmate quality. To measure peer quality, the average achievement level of a student's classmates is computed. To avoid the reflection problem (Manski, 1993), the achievement level of classmates from the prior year is used. For a measure of teacher quality, teacher value-added for the year before the student is assigned to the class is computed.[23]

Figure 6 shows the reduced-form relationship between the classmate quality and teacher quality and the number of students in that subgroup. Visually, there is a large increase in both teacher quality and average classmate achievement (in math) once the school has more than forty students in that subgroup. Table 3 reports the estimates underlying these figures: teacher value-added increases by about 0.06-0.07σ, which is statistically significant. Average classmate achievement shows a similar large increase, with classmates of students facing subgroup-specific accountability scoring about 0.04σ higher on the prior year's math

related work.

[22]The differences between grades 2 and 3 and the rest of the grades are statistically significant at the ten percent level with covariates.

[23]Teacher value-added is computed using a full set of controls including cubic controls for a students' prior test score in both math and English. The value-added is measured in math scores.

tests. The increase, however, is not statistically significant.

5 Discussion

The results in Section 4 indicate that students in a subgroup facing accountability pressure receive a large boost to their test scores. Thus it appears that the subgroup mandates of NCLB achieved their goal of reducing the achievement gaps between minority and non-minority students. However, evidence suggests that these improvements are driven by changes in student's classroom assignments: students that are from a subgroup facing accountability pressure are placed into classes with high-achieving classmates and better teachers. Given this, it is possible that the improvement in these students' outcomes are occurring to the detriment of students who are not in that subgroup.

Looking at students that are not within the subgroup facing accountability pressure, I do not find any negative effects of the change in classroom assignments. In light of the shift in school resources directed to student in the accountable subgroup, one may expect to see negative effects on the other students as they are assigned classes with lower achieving peers or worse teachers.[24] However, it should be noted that the regression discontinuity design may struggle to identify effects on these students due to a lack of statistical power. Specifically, around the threshold there are forty students in the subgroup, but hundreds of students that are not a part of the subgroup. Detecting effects among such a large group where only a few students are likely affected through worse classroom assignments requires substantial statistical power.

The results further show that schools use tracking and matching in order to improve test scores of certain students. Based on prior literature, school administrators likely track and match students based, at least in some part, on parental pressure for better peers and teachers or teacher pressure for less disruptive students (Kalogrides and Loeb, 2013). Under

[24]Or, if the subgroup student is assigned to the better class with no corresponding reassignment, face a higher average class size.

accountability schemes such as NCLB that places higher value on certain types of students, however, schools alter their tracking and matching procedures to the benefit of those students. Policymakers can therefore attempt to reduce achievement gaps by influencing the manner in which school administrators assign students to classes. In fact, if teacher and peer effects are non-linear,[25] there may even be an opportunity to improve overall student achievement by altering these classroom assignments, something that will hopefully be answered through a more structural approach.

A key method of variation is currently not being used in this paper: the dynamic gaming that causes schools to focus their efforts on the earliest grades in a proficiency-based scheme. In Table 2, it was clear that the estimated effects are much larger for grades 2 and 3 students relative to students from other grades. In turn, this implies that a K-5 school focuses its effort for differential tracking and matching on these grades, with no major changes being made to grade 4 students who will only take one more test at the school. This provides another method to investigate the importance of tracking and matching on achievement gaps and student achievement more generally, something that I aim to do in the near future with a difference-in-discontinuity design.

6 Conclusion

Students face unequal levels of school inputs, which lead to persistent achievement gaps among children of different racial and socioeconomic backgrounds. Policymakers have attempted to deal with these gaps by explicitly incorporating achievement gaps into their accountability schemes. Under NCLB, policymakers attempted to reduce these gaps by setting subgroup-specific targets that schools had to attain in order to avoid harsh sanctions.

I shed light on the effect of this policy by using a regression discontinuity design that exploits the rule that the subgroup-specific target only exists if there are forty or more

[25]The most obvious reason that these may be non-linear is that involved parents may react to their kids receiving a worse teacher or worse peers by increasing their educational investment. There is evidence that this occurs in response to class size (BonesrØnning, 2004; Datar and Mason, 2008).

students in that subgroup. I show that schools with the subgroup-specific target have drastic test score improvements among students in that subgroup: about a 0.10σ and 0.02σ increase in math and English test scores, respectively. Subsequently, I uncover the mechanism that schools use to deliver these impressive subgroup improvements: altering their classroom assignments so that these students are in classes with higher achieving peers and better teachers. Policymakers may be able to use these changes in student tracking and matching to develop accountability schemes that change the assignment of students to classrooms and teachers in order to shrink persistent inequality gaps.

While much of the student achievement gaps are driven by the schools that students attend, there is a significant amount of resource inequality within schools. For policymakers, the first type of inequality may be difficult to solve as it is driven by decentralized choices by parents on where to live and whether or not their child attends private school (Nechyba, 2006). Within school inequality, on the other hand, is driven by centralized decision-making by school administrators and thus may be amenable to policy prescriptions. The results in this paper indeed show that changing the incentive structure in a school can changes the within school sorting choices undertaken by school administrators. Further research on how students can be sorted within school to alleviate inequality is sorely needed; and is the future research path of this paper moving forward.

References

Ahn, Thomas and Jacob Vigdor (2014), "The impact of no child left behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina." Working Paper 20511, National Bureau of Economic Research, URL `http://www.nber.org/papers/w20511`.

BonesrØnning, Hans (2004), "The determinants of parental effort in education production: Do parents respond to changes in class size?" *Economics of Education Review*, 23, 1–9.

Carrell, Scott, Richard L. Fullerton, and James West (2009), "Does your cohort matter? Measuring peer effects in college achievement." *Journal of Labor Economics*, 27, 439–464.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a), "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review*, 104, 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b), "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review*, 104, 2633–2679.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006), "Teacher-student matching and the assessment of teacher effectiveness." *Journal of human Resources*, 41, 778–820.

Datar, Ashlesha and Bryce Mason (2008), "Do reductions in class size crowd out parental investment in education?" *Economics of Education Review*, 27, 712–723.

Dieterle, Steven, Cassandra M. Guarino, Mark D. Reckase, and Jeffrey M. Wooldridge (2015), "How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added." *Journal of Policy Analysis and Management*, 34, 32–58.

Finley, Merrilee K. (1984), "Teachers and tracking in a comprehensive high school." *Sociology of Education*, 233–243.

Fryer, Roland G. and Steven D. Levitt (2004), "Understanding the black-white test score gap in the first two years of school." *Review of Economics and Statistics*, 86, 447–464.

Kalogrides, Demetra and Susanna Loeb (2013), "Different teachers, different peers the magnitude of student sorting within schools." *Educational Researcher*, 42, 304–316.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger (2013), "Have we identified effective teachers? Validating measures of effective teaching using random assignment." In *Research Paper. MET Project. Bill & Melinda Gates Foundation*, Citeseer.

Koedel, Cory and Julian R. Betts (2011), "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique." *Education Finance and Policy*, 6, 18–42.

Lareau, Annette (2000), *Home advantage: Social class and parental intervention in elementary education*. Rowman & Littlefield Publishers.

Macartney, Hugh (2016), "The dynamic effects of educational accountability." *Journal of Labor Economics*, 34, 1–28.

Manski, Charles F (1993), "Identification of endogenous social effects: The reflection problem." *Review of economic studies*, 60, 531–542.

McCrary, Justin (2008), "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics*, 142, 698–714.

Neal, Derek and Diane Whitmore Schanzenbach (2010), "Left behind by design: Proficiency counts and test-based accountability." *Review of Economics and Statistics*, 92, 263–283.

Nechyba, Thomas J. (2006), "Income and peer quality sorting in public and private schools." *Handbook of the Economics of Education*, 2, 1327–1368.

No Child Left Behind Act of 2001 (2002), "Pub. L. 107-110. 115 Stat.1440. 8 Jan 2002." URL `http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf`.

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

Rothstein, Jesse (2010), "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics*, 125, 175–214.

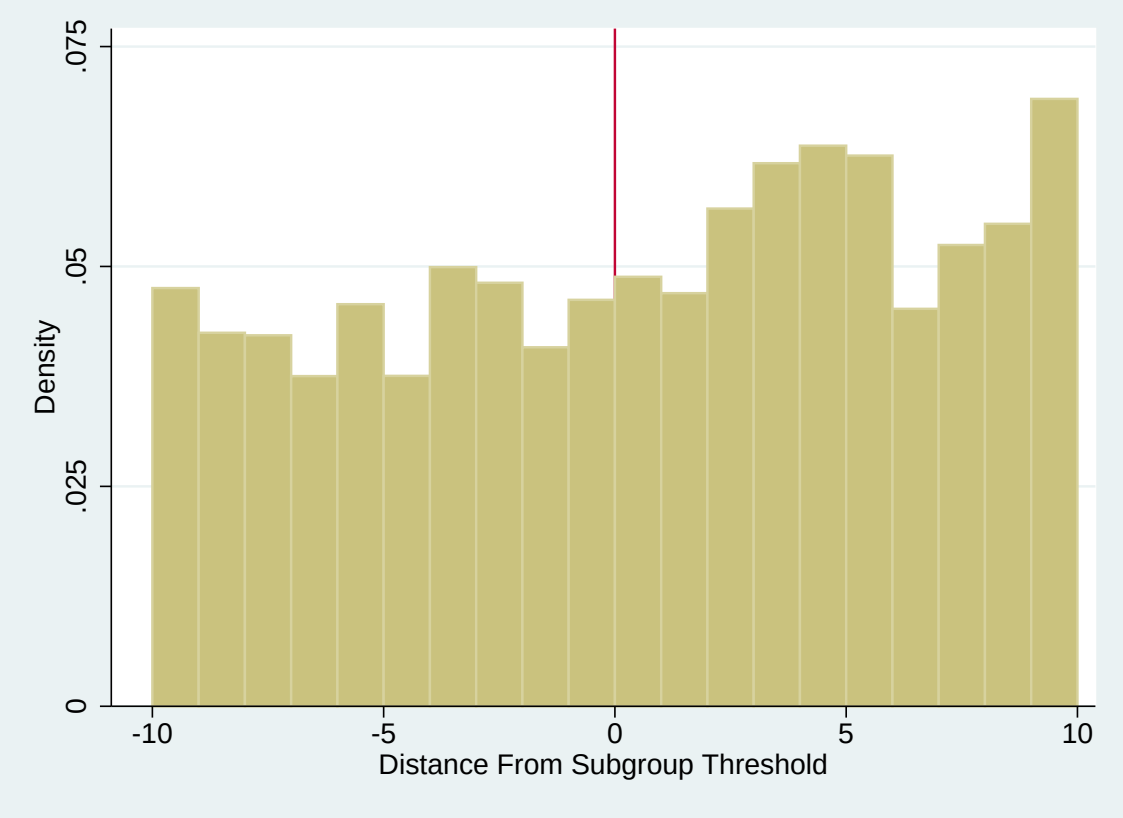
Sacerdote, Bruce (2001), "Peer effects with random assignment: Results for Dartmouth roommates." *Quarterly Journal of Economics*, 116, 681–704.

US Department of Education (2009), "Race to the top program: Executive summary."

Zimmerman, David J. (2003), "Peer effects in academic outcomes: Evidence from a natural experiment." *Review of Economics and statistics*, 85, 9–23.

Figure 1: Density of the Running Variable

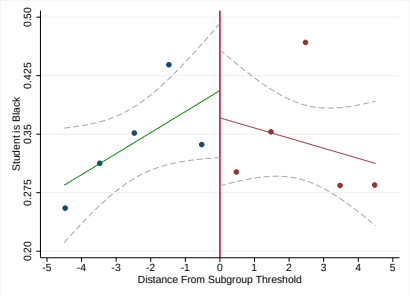
(a) Distance from Subgroup Threshold
McCrary p-value: 0.93



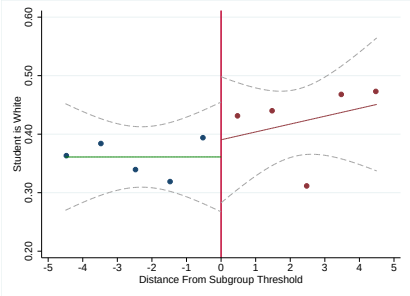
Notes: Figure 1 is based on 55,275 observations. The vertical line indicates the forty student threshold. The p-value from the McCrary (2008) test is computed using a bandwidth of 5 and a bin width of 1. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Figure 2: Covariates

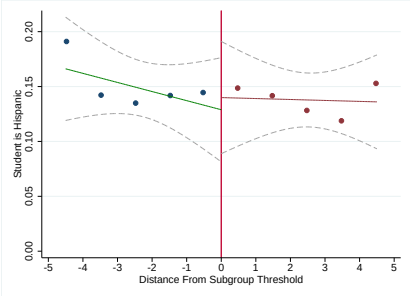
(a) Student is Black
RD Estimate: -3.50 (6.10)



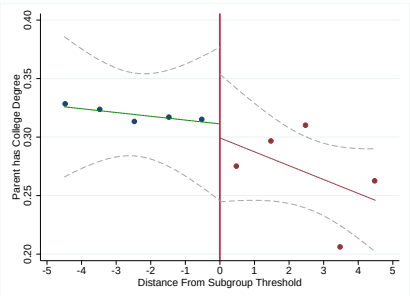
(b) Student is White
RD Estimate: 2.93 (7.35)



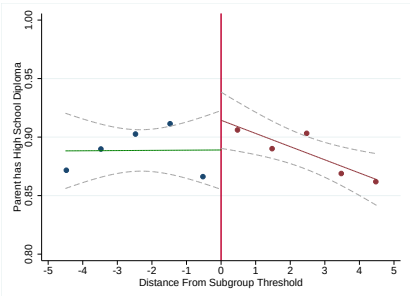
(c) Student is Hispanic
RD Estimate: 1.11 (3.52)



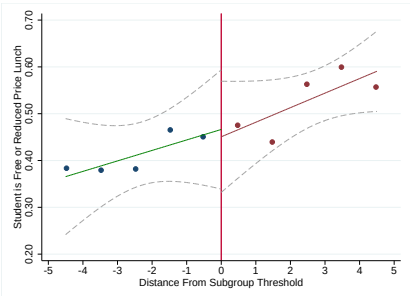
(d) Student has Parent with College Degree
RD Estimate: -1.20 (4.14)



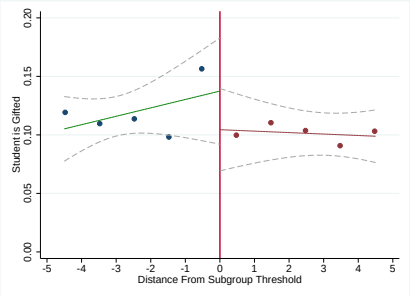
(e) Student has Parent with High School Diploma
RD Estimate: 2.54 (2.15)



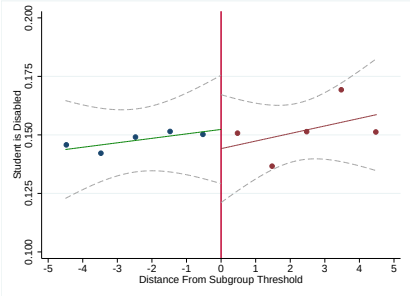
(f) Student is Free or Reduced Price Lunch
RD Estimate: -1.56 (8.01)



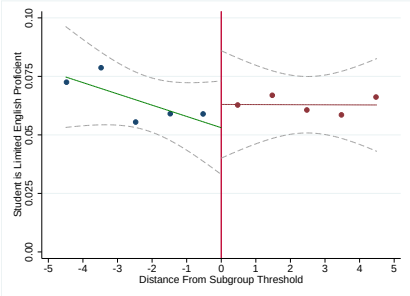
(g) Student is Gifted
RD Estimate: -3.30 (2.89)



(h) Student has a Disability
RD Estimate: -0.82 (1.53)



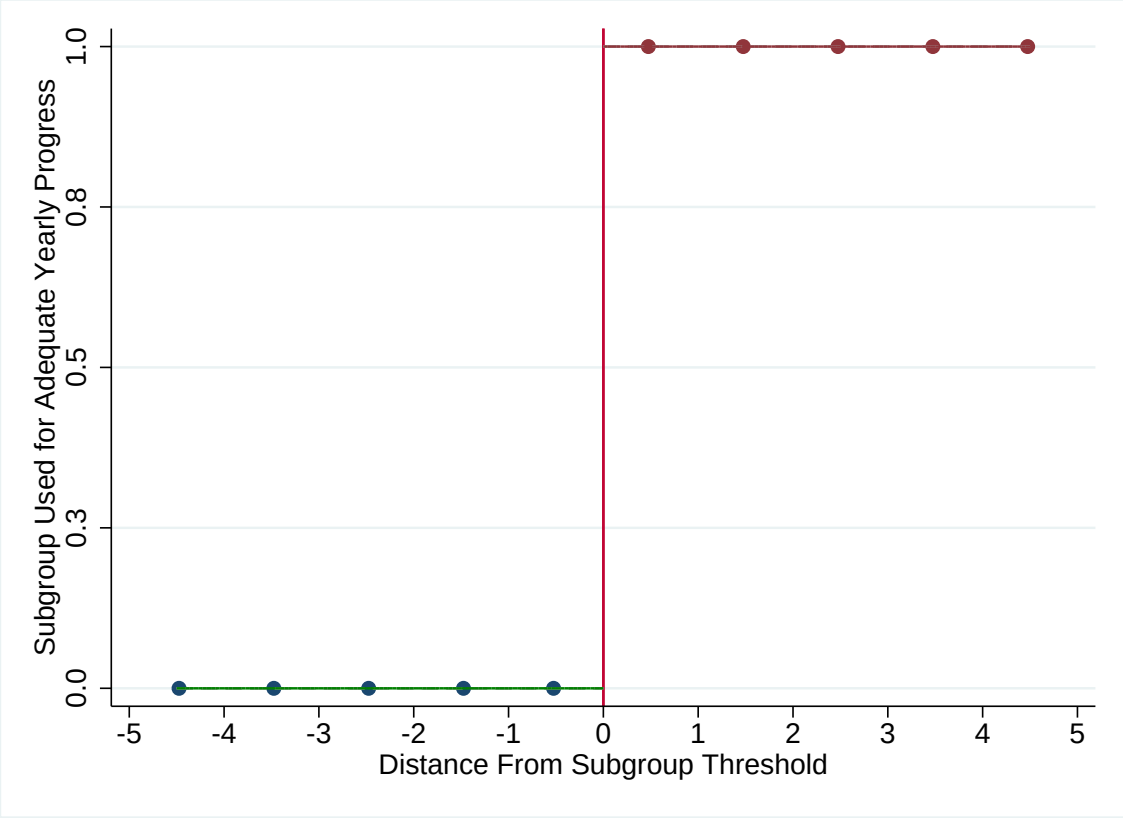
(i) Student is English Learner
RD Estimate: 0.99 (1.48)



Notes: Figures are based on 19,107 observations. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Figure 3: First-Stage (Faced Subgroup Accountability Pressure)

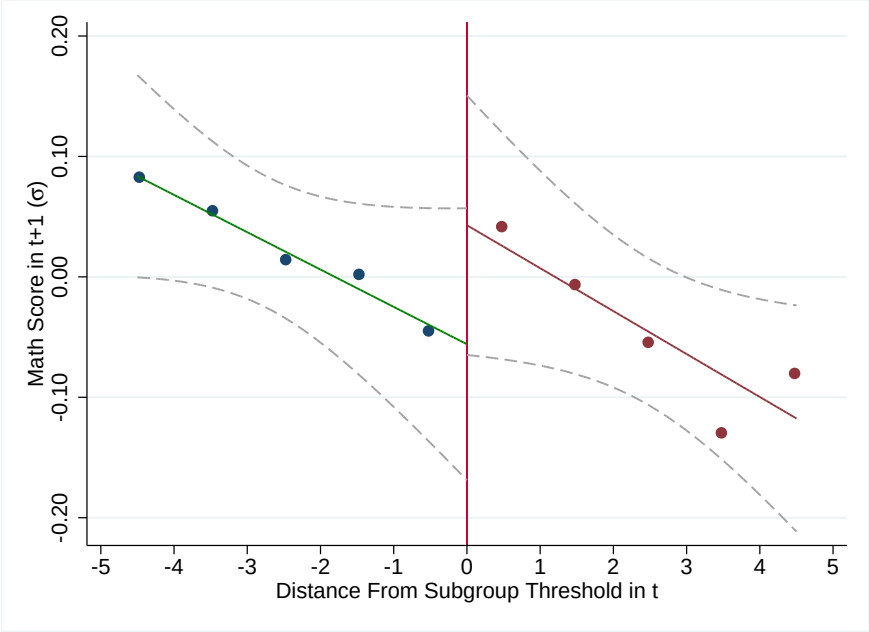
(a) Had Subgroup AYP Target
RD Estimate: 1.00*** (-)



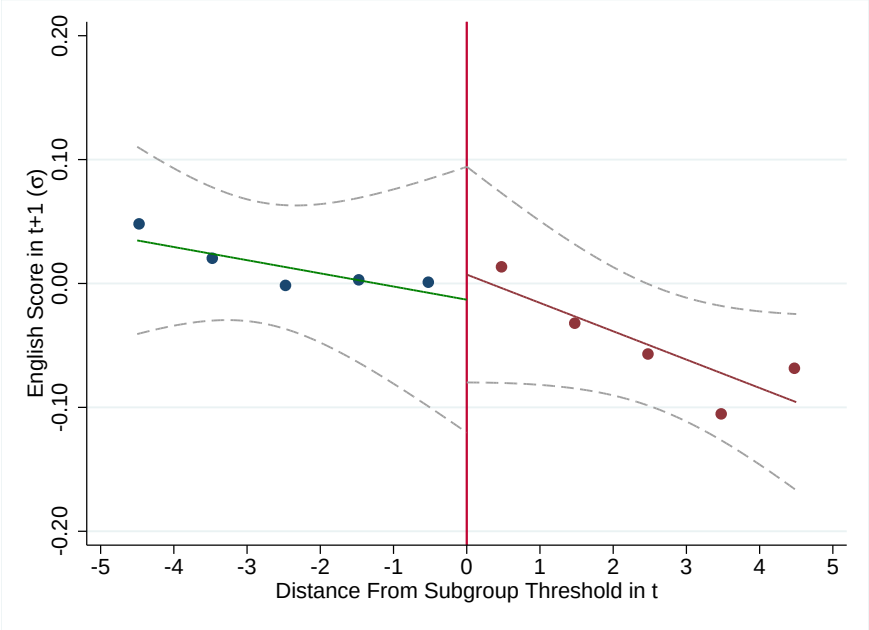
Notes: Figure 3 is based on 22,483 observations. The RD estimate has no standard errors as the line is perfectly fitted. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Figure 4: Reduced-Form

(a) Math Score
RD Estimate: 0.099 (0.076)



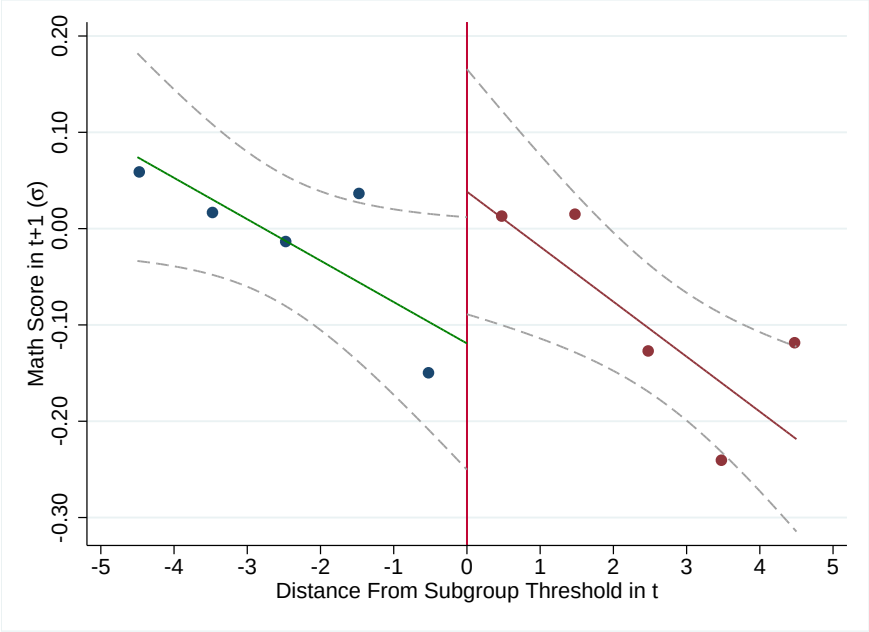
(b) English Score
RD Estimate: 0.020 (0.067)



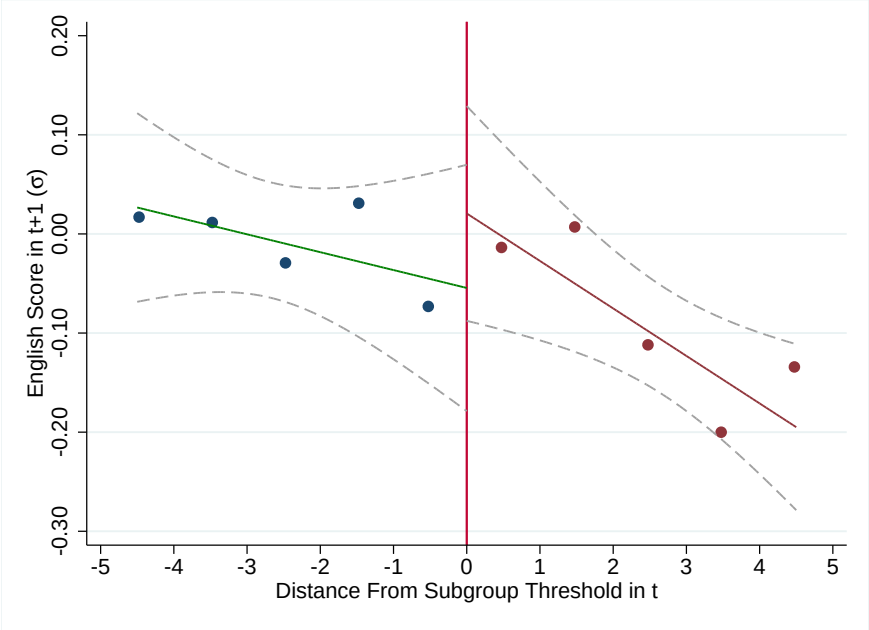
Notes: Figure 4(a) and 4(b) are based on 19,107 and 19,026 observations, respectively. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level. Point estimates correspond to those in panel A column (1) and (3) of Table 2. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Figure 5: Reduced-Form, Grades 2 & 3

(a) Math Score
RD Estimate: 0.158* (0.094)



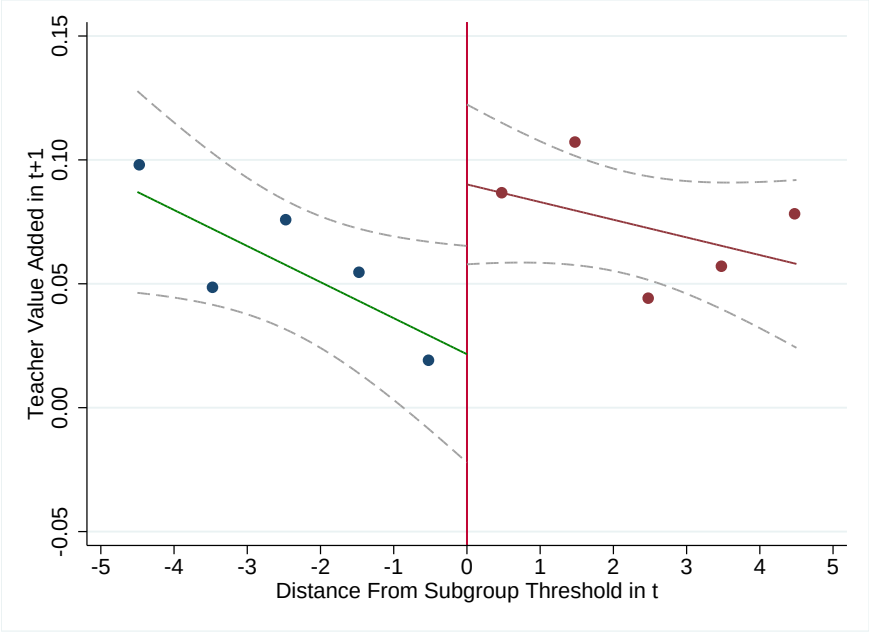
(b) English Score
RD Estimate: 0.075 (0.084)



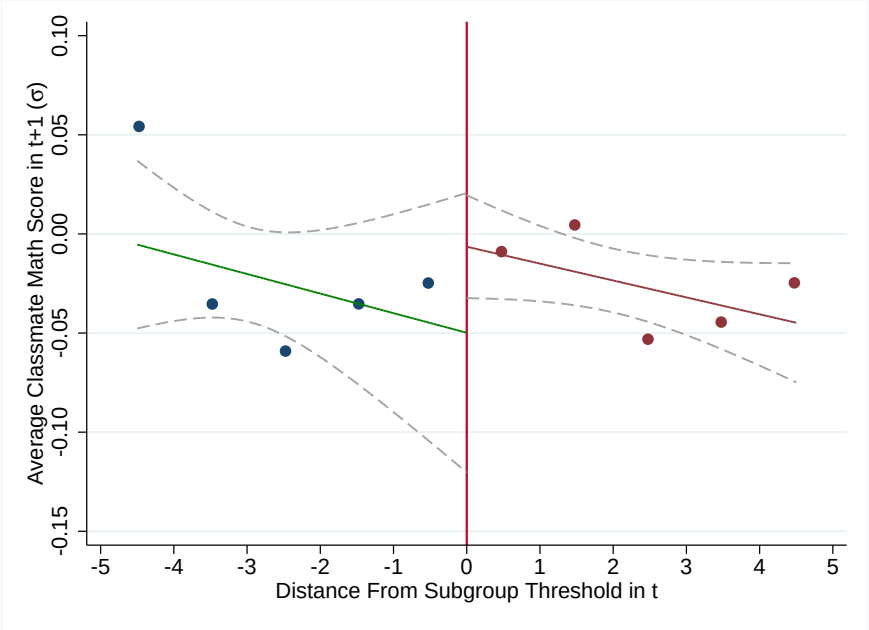
Notes: Figure 5(a) and 5(b) are based on 7,357 and 7,306 observations, respectively. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level. Point estimates correspond to those in panel B, column (1) and (3) of Table 2. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Figure 6: School Inputs, Grades 2 & 3

(a) Teacher Value-added
RD Estimate: 0.069*** (0.027)



(b) Average Classmate Achievement (Math)
RD Estimate: 0.043 (0.038)



Notes: Figure 6(a) and 6(b) are based on 7,344 and 7,651 observations, respectively. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level. Point estimates correspond to those in column (1) of Table 3. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Table 1: Summary Statistics

	All Students [1] (1)	RD Sample (+/- 5)[2] (2)	RD Grade 2 & 3 Sample (+/- 5)[2] (3)
Mean (S.D.) Among Student Characteristics			
Math Score (σ)	0.00	-0.01	-0.07
	(1.00)	(0.97)	(0.95)
English Score (σ)	0.00	-0.01	-0.06
	(1.00)	(0.96)	(0.97)
% White	57.48	38.38	41.95
	(49.44)	(48.63)	(49.35)
% Black	29.76	32.89	37.04
	(45.72)	(46.98)	(48.29)
% Hispanic	7.03	15.48	11.66
	(25.56)	(36.18)	(32.10)
% Asian	1.99	7.88	4.64
	(13.95)	(26.94)	(21.04)
% Parents with High School Diploma	89.88	88.16	87.02
	(30.15)	(32.31)	(33.61)
% Parents with College Degree	31.86	29.06	25.47
	(46.60)	(45.40)	(43.57)
% Free or Reduced Price Lunch	46.73	46.66	57.20
	(49.89)	(49.89)	(49.48)
% Gifted	13.24	10.57	6.70
	(33.89)	(30.75)	(25.00)
% of Students With Disability	14.12	16.80	18.27
	(34.82)	(37.39)	(38.64)
% English Learners	4.25	7.35	7.05
	(20.18)	(26.10)	(25.60)
Observations	2,088,662	27,770	10,342

[1] All students include all grade 3-8 students in a public school in North Carolina in the 2002-03 through 2004-05 school years.

[2] The sample is restricted to students in a subgroup that is +/- 5 students away from the forty student threshold.

Table 2: Regression Discontinuity Estimates of Student Achievement

	Math Scores (σ)		*English Scores (σ)*	
	No Covariates	Covariates	No Covariates	Covariates
	(1)	(2)	(3)	(4)
Panel A. All Grades				
τ_{SRD}	0.099	0.056*	0.020	-0.002
	(0.076)	(0.032)	(0.067)	(0.030)
Panel B. Grades 2 and 3				
τ_{SRD}	0.158*	0.104**	0.075	0.021
	(0.094)	(0.046)	(0.084)	(0.051)
Observations	19,107	18,943	19,026	18,862

Notes: There are 7,357 and 7,306 students in the the grade 2 and 3 math and English samples, respectively. Effect sizes are in standard deviations in the distribution of school-grade means. The bandwidth used is five. Covariates include student level controls for gender, ethnicity, exceptionality status, English learner status, free lunch status, parental education, and grade, year and district fixed-effects. Standard errors are clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 3: Regression Discontinuity Estimates of School Inputs

	Next Year	
	No Covariates	Covariates
	(1)	(2)
Panel A. Teacher Value Added (Math)		
τ_{SRD}	0.069***	0.058**
	(0.027)	(0.025)
Panel B. Average Classmate Achievement		
τ_{SRD}	0.043	0.036
	(0.038)	(0.037)
Observations	7,344	7,312

Notes: Both measures are in math standard deviations. The average classmate achievement measure omits the student's achievement level in the calculation. The bandwidth used is five. Covariates include student level controls for gender, ethnicity, exceptionality status, English learner status, free lunch status, parental education, and grade, year and district fixed-effects. Standard errors are clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.