School Accountability and the Dynamics of Human Capital Formation

Michael Gilraine*
Department of Economics
New York University

November 1, 2017

ABSTRACT

This paper sets out a new approach that enables me to credibly identify dynamic interactions among school inputs for the first time. To do so, I use detailed administrative data from North Carolina to take advantage of an understudied feature of No Child Left Behind – the largest accountability scheme ever implemented in the United States – whereby schools are effectively held accountable only if there are forty or more students in their demographic group. This variation allows me to compare the achievement of students who face accountability to those who do not using a regression discontinuity (RD) design. I then develop a new identification strategy that incorporates year-to-year treatment variation in the RD design to provide the period-by-period randomization necessary to identify interactions in school inputs across time. I find complementarities among inputs across time: RD estimates show that accountability in the next year leads to a 0.18σ test score increase among those receiving treatment in the prior period (relative to those who did not). The reduced-form responses suggest that educators have a sense of the production technology, providing an opportunity to identify the technology structurally. In particular, the reduced-form estimates capture school responses, which are then rationalized by a theoretical model as the relative benefit of investment in each period, pinning down any complementarities in inputs across time. With estimates of the dynamic technology in hand, I consider the efficacy of alternative accountability schemes: conditioning on *initial* test scores rather than prior test scores can both increase average achievement and reduce the pervasive test score gaps that plague public education today.

Keywords: Human Capital; Education Production; Dynamic Complementarities; Dynamic Incentives; School Accountability; Regression Discontinuity; Value-added; Counterfactual Analysis.
***JEL* codes:** C32, I21, I24, I28.

*Thanks to Robert McMillan for helpful discussions about this topic. Thanks also to Natalie Bau, Philip Oreopoulos, Rahul Deb, Aloysius Siow, Debraj Ray, Adam Lavecchia, Uros Petronijevic, Mathieu Marcoux, Jessica Burley and seminar participants at Columbia University, New York University, Northwestern University, Queen's University, UCLA, University of Ottawa, University of Toronto, and the University of Wisconsin-Madison for additional comments. All remaining errors are my own. Contact: Department of Economics, New York University, 19 West 4th Street, New York, NY 10012. Email: mike.gilraine@nyu.edu.

1 Introduction

When studying education policy, it is natural to adopt a dynamic viewpoint, given that the learning process is inherently cumulative. Yet, such a perspective is rarely adopted in empirical research. In no small part, this is due to the stringent requirements associated with uncovering the underlying education production technology. For instance, identifying interactions among inputs across time in an observational setting requires period-by-period randomization – something tantamount to having "lightning [...] strike twice," in the apt phrase of Almond and Mazumder (2013).

I develop a new dynamic approach that allows me to credibly identify dynamic interactions among school inputs for the first time.[1] To do so, I adapt the RD methodology to take advantage of the dynamic structure inherent in many RD applications, whereby entities receive shocks to their forcing variable, generating effective period-by-period randomization. With multi-period randomization in hand, I test for dynamic complementarities in the learning technology explicitly by examining whether treatment effects depend on the prior period's treatment status.

I apply my methodology in the context of the largest accountability reform ever implemented in the United States: the No Child Left Behind Act of 2001 (NCLB). NCLB holds schools accountable for student achievement levels both overall and in nine student subgroups, several of which are based on race. Schools are only held accountable for a given subgroup in a given year, however, if there are forty or more students in the subgroup in that year. I begin by using the subgroup rule in a regression discontinuity (RD) design to calculate the reduced-form policy effects of NCLB accountability on student achievement. Here, I find that schools facing subgroup-specific accountability pressure show large improve-

[1] Heckman, Moon, Pinto, Savelyev, and Yavitz (2010), Aizer and Cunha (2012), Malamud, Pop-Eleches, and Urquiola (2016), and Johnson and Jackson (2017) explore dynamic complementarities between early investments (such as preschool) and later investments, while Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) use a structural model to identify complementarities among parental inputs. My focus is on credibly estimating dynamic complementarities among *school* inputs, and exploring their implications for school behaviour and incentive design.

ments for students in that subgroup, with those students receiving a 0.05σ and 0.03σ boost to their math and reading scores, respectively. I also show that there is no discernible effect of subgroup-specific accountability for students who do not belong to the accountable subgroup.

Next, I take advantage of the dynamic structure of the RD design, where schools are effectively randomized period-by-period around the forty student threshold, to condition RD estimates according to whether the school was treated or untreated in the prior period. Multi-dimensional RD estimates show that accountability in the second year leads to a 0.18σ test score increase among students receiving treatment in the prior year, relative to when they did not. Using the fact that first period students were not present at treated schools in the prior period (but their teachers were), I rule out the alternative story that differences in RD estimates by prior treatment status arise from a dynamic complementarity in the teaching technology (rather than in the student learning technology).

While results from the multi-dimensional RD design establish the presence of dynamic complementarities, they do not pin down the parameters of the student learning technology. In particular, if students are treated in the prior period, educators will raise their investments today because dynamic complementarities increase the marginal productivity of investment. Indeed, the large treatments effects among students that were treated in the prior year (relative to those that were not) indicate that educators know (or gain a sense of) that technology, as these differences would otherwise imply an implausibly large dynamic complementarity parameter. Since the differential RD estimates by prior treatment status incorporate the effects of both the technology and educator responses, a more nuanced approach is required.

If educators know (or gain a sense) of the technology, then I can use educators' responses to NCLB to estimate the parameters structurally. As a starting point, I set out a theoretical model to explore the way a dynamic human capital formation technology affects school responses under fixed-target accountability schemes such as NCLB – those that set a common achievement target for all students to attain. The model generates new insights into the

dynamic incentives of schools: specifically, I show that schools are incentivized to target students in early grades and with ability *below* the achievement threshold, in order to raise those students gradually toward the target. Further, dynamic complementarities sharpen these dynamic incentives since fixed-target schemes allow schools to internalize the benefits of any interactions among school inputs across time. In contrast, models featuring static technologies typically imply that schools will target students near the proficiency threshold[2] and cannot generate the predictions of the dynamic model without contradicting important findings in the empirical literature.[3]

In line with the theoretical model, I find evidence of large accountability-induced test score increases for low-ability students in early grades. For instance, I estimate that subgroup accountability raises math scores for grade 3 students with ability levels below the achievement target by 0.12σ, which is significantly larger than for any other grade or ability group. To interpret these findings, it is useful to consider two types of alternative model – those with myopic educators (and schools), and without dynamic complementarities, respectively. Models with myopic educators would predict that treatment effects should be most pronounced for students around the achievement threshold, rather than for students well-below it: models without dynamic complementarities would imply that treatment effects are independent of prior-year treatment. Against these, the reduced-form evidence points to the existence of both forward-looking educators and dynamic complementarities (whereby the marginal productivities of school inputs are higher in conjunction with prior inputs).

Given the incentives in place, I then use educators' responses to NCLB to estimate the parameters of the technology structurally, and in a credible way. In particular, the period-specific reduced-form estimates that condition on prior treatment status capture school responses to accountability. Assuming (as the reduced-form results suggest) that the school knows the underlying technology, the theoretical model rationalizes the reduced-form esti-

[2]See, for instance, Reback (2008) and Neal and Schanzenbach (2010).

[3]Notably, Cunha and Heckman (2008) find that the marginal return to investment is higher in younger children.

mates under the known NCLB incentive scheme, allowing those responses to be interpreted as the relative benefit of student-specific changes to school inputs in each period, which pins down any dynamic complementarities.[4] In turn, the parameter that determines the relative costs and benefits of inputs chosen by the school is determined, given the complementarity parameter, from the first-order conditions of the school's problem. The structural results underline the importance of dynamic complementarities: they account for about fifty percent of the improvement in student achievement under NCLB (relative to no accountability scheme).

Next, given the dynamic linkages in school inputs across time, I develop a framework for dynamic counterfactual analysis that builds on a general approach in Macartney, McMillan, and Petronijevic (2015). The framework uses the known incentives of various accountability schemes to identify dynamic school responses under those schemes, drawing on the theoretical model and the structural parameters. Using the dynamic framework, I can then simulate the full distribution of student achievement under alternative accountability reforms, including value-added schemes, which set student-specific targets based on prior test scores.

The counterfactual analysis generates several new insights. First, I propose a 'multiperiod value-added' scheme that can counteract the well-known dynamic disincentive in value-added schemes whereby schools avoid investing early in students since those investments create higher achievement targets for the school to attain in future periods (see Macartney (2012)). To avoid this type of dynamic disincentive, the multiperiod value-added scheme sets student-specific targets based on test scores attained when the student *entered* the school. Second, I quantify the impact of eliminating these dynamic disincentives: test score targets set according to *baseline* test scores, as in the multiperiod value-added scheme, raise student achievement by 0.18σ relative to a traditional value-added scheme. When deciding between fixed-target and multiperiod value-added schemes, however, policymakers face a trade-off

[4]To interpret the reduced-form responses as the relative benefit of student-specific changes to school inputs in each period, I assume a similar functional form for the production technology to that in Aizer and Cunha (2012).

between average student achievement and inequality.[5] For example, replacing NCLB with a multiperiod value-added scheme increases average test scores by 0.25σ, but also leads to a twenty percent increase in the black-white test score gap. The test score gap widens under value-added schemes because they motivate schools to focus on all students, while schools under NCLB target low-ability students. Third, the newly-uncovered structural parameters allow me consider alternative schemes that both increase average achievement and *reduce inequality*: in particular, altering the benefit schedule by raising the reward a school receives when a black student (relative to another student) exceeds the threshold in the multiperiod value-added scheme delivers the same black-white test score gap as NCLB, but where average achievement is 0.24σ higher.

The methodology used in this paper provides a springboard for researchers and policymakers to understand the underlying human capital production technology in a variety of other settings. In the context of estimating the effect of class size on student achievement, for example, Krueger (2003) stated: "[t]here is no substitute for understanding the specifications underlying the literature" – a fitting quote in an area where researchers have reached conflicting conclusions, potentially because alternative assumptions about the technology have led them to estimate different empirical specifications. I highlight the importance of estimating this technology in the context of incentive design in education – a particularly relevant topic in light of the recent passing of the Every Student Succeeds Act, which allows states to replace the fixed-target NCLB scheme with value-added schemes.[6] The results imply that using these value-added schemes will increase inequality and may potentially lower average achievement. Yet, the insights in this paper also show how policymakers can construct multiperiod value-added schemes to counteract the dynamic disincentives of traditional value-added reforms and help reduce the pervasive test score gaps that plague public education today.

[5] A similar trade-off in a static setting is highlighted in Macartney et al. (2015).

[6] The Every Student Succeeds Act still requires a fixed-target scheme to be a component of a state's accountability system. The law, however, allows states the flexibility to incorporate additional components to their accountability scheme, including value-added targets.

The rest of the paper is organized as follows: The next section provides background to NCLB and places the analysis alongside the associated literature. Section 3 introduces the theoretical model, which predicts school responses to fixed-target accountability schemes under a dynamic human capital formation technology. In Section 4, I provide an empirical framework to test the model predictions and to determine whether there are dynamic complementarities in school inputs; I also discuss the data set. I present reduced-form results in Section 5, and use these to identify accountability responses to the model structurally in Section 6. In Section 7, I perform counterfactual simulations, and discuss the results in Section 8. Section 9 concludes, drawing out the broader significance of the research.

2 Background and Literature

The No Child Left Behind Act of 2001 (NCLB) – the major education initiative of the Bush administration – aimed to raise educational achievement and to "clos[e] the achievement gap between high- and low-performing 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). In pursuit of that goal, NCLB incentivized schools to improve student achievement by enforcing progressively harsher corrective action, including school takeovers,[7] when a school failed to meet Adequate Yearly Progress (AYP). AYP was determined by the proportion of students whose test scores fell below a pre-set 'proficiency target,' based on whether a student attained a pre-defined score on the end-of-grade standardized math and reading tests. If the proportion of proficient students in a school was below the state target, in either math or reading, it failed AYP.[8] Test results for grades 3-8 were used for NCLB purposes in elementary and middle schools.

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

[8]There were some exceptions written into the law such 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 (see Spellings (2005) for more information). In addition, there are additional targets that the school also must meet to pass AYP: more than 95 percent of students overall and in each subgroup have to take the end-of-grade tests and the school is required to have 90 percent attendance. In North Carolina, less than two percent of schools in my sample fail one of these requirements

A major feature of NCLB involves the subgroup rules that are embedded in the law. According to these, a school has to reach its proficiency target both overall and in each of nine subgroups: black, Hispanic, white, Asian, multi-racial, native American, economically disadvantaged, limited English-proficient and students with disabilities. For example, in 2003, a school in North Carolina passed AYP if at least 74.6 percent of their student body and 74.6 percent of students in each of the nine subgroups were labelled proficient. Since one of the subgroups in a school is often lower-performing than the school as a whole, many schools reach the overall target, but fail AYP due to the subgroup targets.[9]

Schools may have few students in a given subgroup, which could make their ability to pass the subgroup targets hinge on a small number of students. Thus, legislators incorporated a rule in the law that there must be a minimum number of students in a given subgroup for the school to be held accountable for that subgroup's proficiency target. The threshold number of students varied by state: in North Carolina, which is the focus of my analysis, the threshold is forty students. Therefore, any school in North Carolina with forty or more students in a subgroup faced accountability pressure for that subgroup, while schools with less than forty students did not. In Section 4, this rule provides identifying variation for the methodology I develop to investigate the causal effects of NCLB on human capital formation.

Treatment Status: The count of the number of students in a subgroup only includes students in grades 3 and above since K-2 students are not used to calculate AYP determinations. These counts are complicated by NCLB rules whereby students are used for accountability purposes only if they are present for seventy-five percent of school days (about 140/180 school days). Accordingly, schools often had fewer than forty students in a subgroup take the end-of-grade test but were held accountable for that subgroup, and vice versa. For example, a school that had thirty-nine disadvantaged students in September would be held accountable for the disadvantaged subgroup if a disadvantaged student entered their school in mid-October, but not if that student arrived in mid-November. Thus, schools with

[9]Stullich et al. (2006) show that among schools failing AYP, 41 percent passed the overall proficiency target but failed AYP as one or more subgroups did not reach their proficiency target.

around forty students in a subgroup were uncertain whether that subgroup would be held accountable under NCLB until several months into the school year.

The uncertainty over treatment status for schools around the forty student threshold may dampen school responses in the empirical results, attenuating the reduced-form results in Section 5 towards zero. For the structural estimation exercise, underestimation of reduced-form parameters may lead to some underestimation of the structural parameters, though this is unlikely to be a serious concern given that the parameters are identified primarily by the ratio of the reduced-form estimates.[10]

Sorting: Because classroom assignments are generally set at the start of the school year, when treatment status is uncertain, the ability of schools to sort students differentially across classes and teachers by subgroup accountability status near the forty student threshold is limited. Thus, the mechanism generating the effects of accountability on student achievement in this paper are unlikely to be caused by reallocating students to better teachers or peers, which Section 8.2 verifies. This indicates that, to a first-order approximation, it is reasonable to abstract from student sorting in the counterfactual analysis. In addition, the absence of sorting aligns with the fundamental motivation behind accountability schemes: accountability raises educator effort, rather than reallocating resources across students. Section 8.2 presents evidence supporting the view that increases in teacher effort are generating the improvements in student achievement.

2.1 Literature

Understanding the full effects of accountability schemes requires knowledge of the underlying human capital production technology. Particular assumptions about the production technology imply distinct empirical specifications even with a given data set, which may in turn generate widely differing conclusions (Todd and Wolpin, 2003). This has been an issue in several literatures, notably class size reduction – see Krueger (1998, 2003). Given their

[10]This idea is elaborated on in Section 6, which notes that a ten percent underestimation of the RD estimates would lead to a five percent rise in the structural parameters.

appeal, recent state-of-the-art papers have begun formulating and estimating *cumulative* production functions (Todd and Wolpin, 2007; Cunha and Heckman, 2008; Cunha, Heckman, and Schennach, 2010), emphasizing the role of parental inputs and dynamic complementarities among them. For example, Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) use a structural model to identify the skill formation technology and find evidence of dynamic complementarities among parental inputs.

School inputs have been the focus of a vast empirical literature in education, but the literature has not examined possible dynamic complementarities among them. This is likely due to the serious challenges involved in estimating *school* input complementarities. In an observational setting, identification of any dynamic school input complementarity would require exogenous variation in schooling inputs that affected a set of individuals in one period, and then for the same set of individuals to receive another exogenous shock to the same school input in a second period – a configuration akin to "asking for lightning to strike twice" (Almond and Mazumder, 2013).

In light of these stringent requirements, several papers provide evidence of dynamic complementarities by calling attention to the fact that the effects of early investment are larger among children with a higher initial stock of skills (Heckman et al., 2010; Aizer and Cunha, 2012; Lubotsky and Kaestner, 2016).[11] Malamud, Pop-Eleches, and Urquiola (2016) examine the strength of complementarities between initial endowments and later investments in an appealing way by combining a shock to initial child ability with a later shock to schooling inputs. While they find no direct evidence of dynamic complementarities, the authors uncover suggestive evidence that parents and children behave in ways that undo such complementarities. Johnson and Jackson (2017) also use two shocks by combining the roll-out of the Head Start preschool program with court-induced changes to K-12 school spending and find that the benefits of Head Start were larger when followed by access to better-funded public schools, indicative of dynamic complementarities. I add to this literature (which

[11]Several papers also find that the effects of interventions are larger among younger children (Cunha and Heckman, 2008; Chetty et al., 2016), which is consistent with dynamic complementarities.

Appendix Section C describes more in-depth) by providing evidence of dynamic complementarities among schooling inputs for the first time, using successive exogenous shocks to school investments.

NCLB has been studied widely, with several careful analyses concluding that NCLB accomplished one of its main goals in that it improved test scores (Jacob, 2005; Dee and Jacob, 2011; Chakrabarti, 2014). Critics have pointed out, however, that NCLB created many perverse incentives, including teaching to the test (Koretz, 2008), manipulating the test-taking pool (Jacob, 2005; Figlio, 2006), or even altering the calorie content of school meals (Figlio and Winicki, 2005).[12] Since school rewards are based on students exceeding a preset achievement target, researchers have also established that schools respond to NCLB by targeting their resources towards students with predicted test scores near the achievement target (Krieg, 2008; Reback, 2008; Springer, 2008; Neal and Schanzenbach, 2010).

The subgroup rules embedded in NCLB have been studied extensively.[13] Prior papers, which generally use difference-in-differences methodologies, conclude that students belonging to subgroups facing accountability pressure receive a boost in their test scores relative to other students in the school (Krieg, 2011; Lauen and Gaddis, 2012; Gaddis and Lauen, 2014).[14] Similar to this paper, Farber (2016) uses Texas's subgroup cutoff rule to estimate the policy effect of subgroup accountability. While my paper reports these policy estimates – roughly similar to those in Farber (2016) – my main focus is on identifying the underlying technology and its importance for the effects, and design, of accountability schemes.

[12]See Figlio and Loeb (2011) for a more in-depth review of gaming under NCLB.

[13]Studies have also examined subgroup rules in other types of accountability schemes. For example, see Deming et al. (2016).

[14]In contrast, Sims (2013) uses the fact that subgroup rules make otherwise-similar schools face different probabilities of failing AYP and finds that the subgroup accountability rule leads to lower future student achievement for all students.

3 Model

This section lays out a theoretical model to investigate school responses to an accountability scheme under a dynamic human capital accumulation technology. In line with the empirical application, I focus on school responses under fixed-target accountability schemes. The theoretical model yields two testable predictions not readily explained by a static model. I test these predictions in Section 5.3. The theoretical model also supplies the backbone for the structural estimation and the dynamic counterfactual framework in Sections 6 and 7, where I quantify the efficacy of alternative accountability schemes.

3.1 Technology

Human capital for student i at time t is given by student achievement, A_{it}, which is a function of the full history of school input spending up to time t, $\boldsymbol{S}_{it}=\{S_{i1},...,S_{it}\}$, student innate ability, α_i, and unobserved random factors, ϵ_{it}. Under a general production function, student i's achievement at any time t is given by:[15]

$$A_{it} = A_t(\alpha_i, S_{i1}, ..., S_{it}, \epsilon_{i1}, ..., \epsilon_{it})\,. \tag{3.1}$$

I impose several plausible restrictions on the general technology. First, the noise in the production technology, ϵ_{it}, is assumed to be unimodal with zero mean, symmetric, exhibiting no serial correlation, and having an additively separable form, with its cdf and pdf given by $H(\cdot)$ and $h(\cdot)$, respectively. Thus,

$$A_{it} = f_t(\alpha_i, S_{i1}, ..., S_{it}) + \epsilon_{it}\,. \tag{3.2}$$

[15]Family inputs are omitted in this model, implying that any parental responses are subsumed in the school inputs (see Todd and Wolpin (2003)). Since schools should take these parental responses into account when making investment decisions, parental responses change the interpretation of S_{it} from solely the effect of school inputs to the combined effect of school inputs and their associated parental responses.

Second, school inputs always affect student achievement non-negatively ($\frac{\partial f_t(\cdot)}{\partial S_{it}} \geq 0 \; \forall i, t$) and the impact of school inputs are decreasing with ability level ($\frac{\partial f_t(\cdot)}{\partial S_{it}\alpha_i} \leq 0 \; \forall t$) – in line with the empirical literature which finds that the marginal productivity of investments is higher for children with lower skills (which is distinct from children with lower prior investments).[16] I also assume that $\frac{\partial f_t(\cdot)}{\partial S_{it}} \geq \frac{\partial f_t(\cdot)}{\partial S_{i,t+1}}$, in line with results from Cunha and Heckman (2008), indicating that investment yields potentially higher returns for younger children.[17] Dynamic complementarities in this production technology imply that school inputs in one period increase the marginal productivity of inputs in later periods, and arise iff $\frac{\partial^2 f_t(\cdot)}{\partial S_{it} \partial S_{i,t-1}} > 0$ (Cunha and Heckman, 2007).

3.2 Incentive Scheme

Each school has a total of N students. Since each student remains in the school for T years,[18] the school maximizes the number of students who exceed the achievement threshold in the current time period and for subsequent time periods $t+1, ..., T$. Under a fixed-target accountability scheme, the policymaker sets an achievement threshold, $\mathcal{A}^*$,[19] and provides rewards to schools with a benefit b for each student who exceeds this threshold. Student ability within the school is uniformly distributed over $[\underline{\alpha}, \bar{\alpha}]$, where $\underline{\alpha} \leq \mathcal{A}^* \leq \bar{\alpha}$. In line with the empirical application, the benefit is only given if a student is held accountable, making benefits both student- and period-specific. Let $\Gamma_{it} = 1$ denote the case where student i is held accountable and $\Gamma_{it} = 0$, the case where student i is not held accountable in period t,

[16] For instance, Angrist et al. (2012) find that the benefits of charter schools are higher among low-ability students. Similarly, Agostinelli and Wiswall (2016) – who use a more structural approach – conclude that the marginal productivity of early investments is higher for children with low existing skills.

[17] This assumption is only required for Proposition 1. Without it, an extremely high marginal productivity of inputs in period $t+1$ relative to period t could make it possible that schools optimally invest such that $S_{i,t+1} > S_{it}$.

[18] In my empirical application, the time horizon varies by student. For convenience, I do not make T student-specific in the theoretical model, and so the model is equivalent to the problem faced by a school for a given student cohort.

[19] For convenience, the threshold is set identically in each grade. If the achievement threshold varies each period, then Proposition 2 requires a "for some t" qualifier to account for a fixed-target scheme with achievement thresholds that vary appreciably across grades. Empirically, the achievement threshold does not vary substantially across grades, although my structural estimation allows the achievement threshold to differ in each period.

where Γ_{it} follows a Bernoulli distribution with T trials and a 'success parameter' p_i, so that $\Gamma_{it} \sim B(T, p_i)$. Benefits from exceeding the achievement threshold are therefore given by:

$$b_{it} = \begin{cases} b & \text{if } \Gamma_{it} = 1 \\ 0 & \text{if } \Gamma_{it} = 0\,. \end{cases} \tag{3.3}$$

In each period, the model timing is such that schools observe the realization of Γ_{it} before choosing the level of inputs for that period, S_{it}. For future time periods $t+1, ..., T$, the school forms an expectation over whether the student will be treated, $\mathbb{E}[\Gamma_{it} = 1]$. Under subgroup accountability, students within a subgroup are either held accountable or not, so I assume that schools form identical expectations over Γ_{it} for these students. I also assume that treatment is not serially correlated (i.e., $P(\Gamma_{i,t+1} = 1|\Gamma_{it} = 1) = P(\Gamma_{i,t+1} = 1|\Gamma_{it} = 0)$) – capturing the essence of the empirical strategy where treatment statuses in periods t and $t+1$ are as good as random. In addition, I assume that schools have rational expectations, so that $\mathbb{E}[\Gamma_{it} = 1] = p_i$, $\forall t$.

Schools face a convex cost of resources, $c(S_{it})$, which makes the costs of inputs student-specific (rather than classroom-specific), in line with evidence provided in Section 8.2, which rules out large within-class spillovers.[20] Under a fixed-target scheme, the school's payoff for each student i is $U_{it} = \Gamma_{it} \cdot b \cdot \mathbb{1}_{A_{it} \geq \mathcal{A}^*} - c(S_{it})$, which in expectation is:

$$U_{it} = p_i \cdot b \cdot H[A_{it} - \mathcal{A}^*] - c(S_{it})\,. \tag{3.4}$$

[20] Abstracting from classroom level effort is relatively harmless here since the empirical application focuses on students in a demographic group with around forty students at the school level. North Carolina schools averaged about five classrooms per grade, implying that most classrooms would only consist of two or three of these students in a K-5 school.

The school's optimization problem in period t can therefore be written as:

$$\begin{aligned} \underset{\{S_{it}\}_{i\in N}}{max}\ & \Gamma_{it}\cdot b\cdot H(A_{it}-\mathcal{A}^*)-c(S_{it})+\sum_{i}^{N}\sum_{t+1}^{T}[p_i\cdot b\cdot H(A_{it}-\mathcal{A}^*)-c(S_{it})] \\ & \textit{subject to:} \qquad S_{it}\geq 0 \qquad \forall i,t\,. \end{aligned} \tag{3.5}$$

The first-order conditions for each i across all time periods are given by:

$$\begin{aligned} &\frac{\partial U_i}{S_{it}}:\Gamma_{it}\cdot b\cdot h[A_{it}-\mathcal{A}^*]\frac{\partial f_t(\cdot)}{\partial S_{it}}+...+p_i\cdot b\cdot h[A_{iT}-\mathcal{A}^*]\frac{\partial f_T(\cdot)}{\partial S_{it}}\leq c'(S_{it}) \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ &\frac{\partial U_i}{S_{iT}}:\Gamma_{it}\cdot b\cdot h[A_{iT}-\mathcal{A}^*]\frac{\partial f_T(\cdot)}{\partial S_{iT}}\leq c'(S_{iT})\,, \end{aligned} \tag{3.6}$$

leading directly to two propositions: schools invest more resources in earlier periods, and they target students with ability below the achievement threshold.

Proposition 1 *If schools invest optimally, $S_{it}\geq S_{i,t+1}\geq ...\geq S_{iT}$, $\forall i$ in expectation.*

Proof: See Appendix A. ■

The intuition behind Proposition 1 is as follows: input levels are higher in earlier periods since an increase in school inputs this year raises the likelihood of a student exceeding the achievement threshold for this year *and* all future years, which the school internalizes when making its input decisions. Therefore, a year t increase in school inputs for student i raises her probability of exceeding the threshold in that year, as well as $t+1$,...,T, while the same increase to school inputs in year $t+1$ only raises her probability of exceeding the threshold in years $t+1$,...,T. Convex costs and dynamic complementarities make it suboptimal to spend all the inputs in the first period.

Corollary *School inputs are higher when schools expect that it is more (rather than less) likely that a student will be held accountable.*

The corollary follows the same logic as Proposition 1: if a school expects that a student

is more likely to be held accountable in the future, then the benefit of raising school inputs this period increases as it raises the likelihood that the school gets a benefit if the student exceeds the achievement threshold. This corollary suggests a direct test of whether schools internalize the future benefits generated by the cumulative nature of the production technology: if schools internalize these benefits, then this indicates that schools have a sense of the underlying technology, providing a foundation for the structural identification of the learning technology through the theoretical model. I directly test this corollary in Subsection 8.1.

Proposition 2 *Within a subgroup, school inputs are highest for some student with innate ability below the achievement threshold.*

Proof: See Appendix A. ■

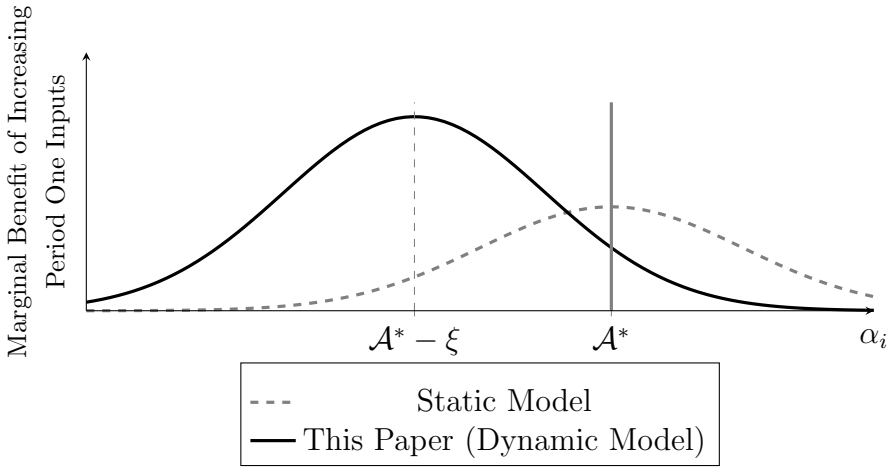
To provide intuition for Proposition 2, Figure 1 provides a stylized representation of the marginal benefit of school inputs as a function of student ability for some period $t < T$. The shape of the marginal benefit profile depends on the distribution of the error terms: a single-peaked distribution (such as normal or type I extreme value) produces an inverse-U shape profile. In addition, I compare the resulting marginal benefit profile to a static model where schools simply maximize the number of proficient students in a given year, as in Neal and Schanzenbach (2010).[21]

In the static model, schools receive the highest benefit by investing in students right around the achievement threshold (in Figure 1(a), $\alpha_i \approx \mathcal{A}^*$). Under a dynamic model, in contrast, the marginal benefit of school inputs is highest among low-ability students who lie well below the achievement threshold. This arises since schools take into account the fact that increases in school inputs this year push students closer to the achievement threshold in future years, when future inputs will propel those students above the threshold. This capacity to raise students gradually toward the achievement threshold moves the marginal benefit profile

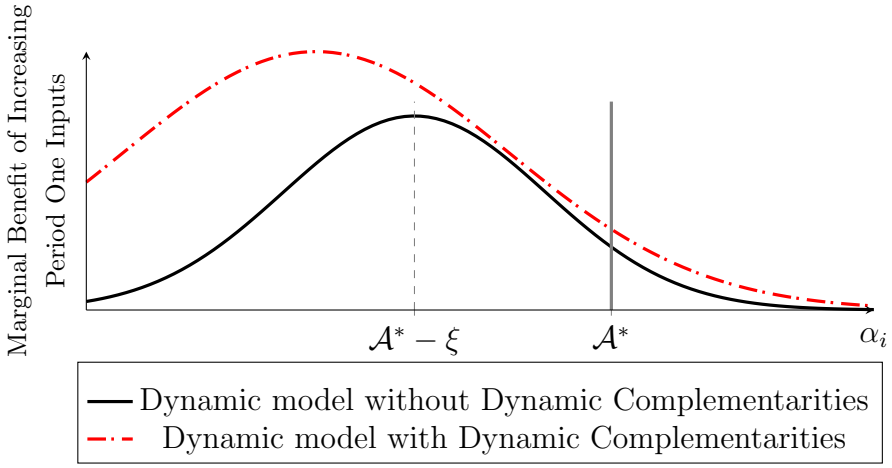
[21]For the static case, the school maximization problem is written as: $\underset{\{S_i\}_{i\in N}}{max} \sum_{i}^{N} [\Gamma_i \cdot b \cdot H(A_i - \mathcal{A}^*) - c(S_i)]$ subject to: $S_i \geq 0 \ \forall i$.

Figure 1: Marginal Benefit of Investment by Ability under a Fixed-target Scheme

(a) Dynamic Relative to Static Model



(b) Adding Dynamic Complementarities



Notes: Figure 1(a) compares a marginal benefit profile of school inputs in my model to that of a model of myopic educators, as in Neal and Schanzenbach (2010). Figure 1(b) then compares the marginal benefit profile in my model with and without dynamic complementarities. The solid vertical line at $\mathcal{A}^*$ represents the achievement target. On the x-axis, ξ represents a constant to show the distance from the achievement target, $\mathcal{A}^*$, in this stylized example.

to the left compared to the static model, such that the marginal benefit is highest for students with ability significantly below the achievement threshold (in Figure 1(a), $\alpha_i \approx \mathcal{A}^* - \xi$, where ξ represents the distance to the achievement threshold). Additionally, since the static model does not take into account how current inputs increase the likelihood of students exceeding the threshold in future periods, the marginal benefit of period t investment is much higher

in the dynamic case (generating the upward shift in the dynamic model). Figure 1(b) shows that dynamic complementarities increase both the marginal benefit of investment and the desire to target low-ability students.

While neither Propositions 1 nor 2 requires dynamic complementarities to hold, the presence of dynamic complementarities strengthen their findings since they raise the marginal benefit of early inputs from the school's perspective. Therefore, if the school knows that the underlying technology features dynamic complementarities, current achievement depends upon prior levels of inputs in two ways: first, directly through the technology, and, second, by the school raising current period inputs due to an increase in the marginal benefit of investment. Consequently, modeling school responses to the underlying production technology is necessary to structurally identify the dynamic complementarity parameter.

4 Empirical Framework

I now describe my empirical approach, built around a regression-discontinuity (RD) design. The RD design provides a reduced-form link between subgroup-specific accountability and student achievement. Furthermore, it generates year-to-year variation in treatment, allowing me to test for dynamic complementarities in the student learning technology directly. The framework will also be used to test model predictions in Section 3.

4.1 Regression Discontinuity

I use a regression discontinuity (RD) design to obtain reduced-form estimates of the effects of accountability on student achievement. The design takes advantage of the rule that schools with more than forty students in a given subgroup are required to meet an achievement target for that subgroup, while schools with less than forty students do not. The essence of the empirical strategy is to compare outcomes in schools with slightly fewer 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. Due to the policy rule, the school with forty disadvantaged students must meet the accountability target for disadvantaged students in order to pass AYP, while the school with thirty-nine disadvantaged students does not face the same requirement. Since schools with forty disadvantaged students are unlikely to differ much from schools with thirty-nine disadvantaged students, we can compare outcomes of disadvantaged students across the two schools to examine the effect of NCLB's subgroup-specific accountability provisions.

As we incorporate schools further away from the forty-student cutoff into the analysis, possible confounding relationships between the number of students in a subgroup and student achievement may appear, making it necessary to control for the number of students belonging to a subgroup in a school through some function, which I assume to be linear in the main analysis.[22] To keep with the spirit of comparing schools close to the cutoff, I restrict attention to only include schools around the forty student cutoff within a bandwidth of five students, though Appendix Figure A3 shows robustness to alternative bandwidths.

Formally, denote a given subgroup by the subscript g. I then regress:

$$\begin{aligned} y_{sgt} = \gamma_1 + \tau_{i \in g} T_{sgt} + \theta_1 X_{sgt} + \phi_1 (X_{sgt} * T_{sgt}) + \lambda_g + \theta_t + \delta_1 Z_{sgt} + \epsilon_{gst}, \\ for \quad w_{sgt} \leq X_{sgt} \leq w_{sgt}, \end{aligned} \tag{4.1}$$

where y_{sgt} represents the average standardized test score of students who belong to subgroup g in school s at time t, X_{sgt} is the number of students in a given subgroup g in school s at time t relative to the forty student threshold (i.e., subgroup enrollment minus forty), T_{sgt} is an indicator equal to one if there are forty or more students in subgroup g in school s at time t (i.e., $T_{sgt} \equiv \mathbb{1}\{X_{sgt} \geq 0\}$), Z_{sgt} is a set of controls, including prior test scores and demographic characteristics, w_{sgt} represents the bandwidth, and λ_g and θ_t are subgroup and year fixed-effects, respectively. The coefficient of interest in Equation 4.1 is $\tau_{i \in g}$, which

[22] Appendix Table A4 examines robustness to alternative control functions.

– under assumptions that are tested in Section 5 – represents the causal effects of NCLB subgroup-specific accountability pressure on student achievement for students who *belong* to subgroup g.

Difference-in-Discontinuity: Spillovers are investigated in the above framework by estimating the effect of subgroup-specific accountability for students who do not belong to the accountable subgroup. Denote whether or not a school-subgroup-year entity consists of students belonging to subgroup g with the subscript b (i.e., $b=1$ if $i \in g$ and zero otherwise). I then estimate the effect of subgroup-accountability on students who belong and who do not belong to the subgroup simultaneously through a difference-in-discontinuity regression:[23]

$$\begin{aligned} y_{sgbt} = \gamma_1 + \tau_{i \notin g} T_{sgt} + \theta_1 X_{sgt} + \phi_1 (X_{sgt} * T_{sgt}) + \lambda_g + \theta_t + \delta_1 Z_{sgbt} \\ + B_{sgbt} [\gamma_2 + \tau_{diff} T_{sgt} + \theta_2 X_{sgt} + \phi_2 (X_{sgt} * T_{sgt}) + \lambda_g + \theta_t + \delta_2 Z_{sgbt}] + \epsilon_{gsbt}, \\ for \quad w_{sgt} \leq X_{sgt} \leq w_{sgt}, \end{aligned} \tag{4.2}$$

where B_{sgbt} is an indicator equal to one if students belong to subgroup g, and all other variables are defined in Equation 4.1. The three coefficients of interest from Equation 4.2 are: $\tau_{i \in g} \equiv \tau_{i \notin g} + \tau_{diff}$ (same parameter as in Equation 4.1), $\tau_{i \notin g}$, and τ_{diff}. $\tau_{i \notin g}$ represents the causal effects of NCLB subgroup-specific accountability pressure on student achievement for students who *do not belong* to subgroup g, while τ_{diff} represents the differential benefit of subgroup accountability accruing to students belonging to an accountable subgroup compared to those who do not. If there are no spillovers from subgroup accountability (i.e. $\tau_{i \notin g} = 0$), then τ_{diff} has the same causal interpretation as $\tau_{i \in g}$. With spillovers, however, τ_{diff} could either under- or over-represent the effect of accountability on students belonging to subgroup g. On one hand, schools may appropriate resources from other students to improve the achievement of students in subgroup g, lowering their achievement; on the other hand, schools may raise their total level of effort or there may be peer effects, benefiting all students. To allow for possible spillovers, the estimates of $\tau_{i \in g}$ are interpreted as the causal

[23] See Grembi et al. (2016) for a description of the difference-in-discontinuity estimator.

effect of accountability in this paper.

Equations 4.1 and 4.2 estimate local average treatment effects (LATE). As already described in Section 2, schools may not know their treatment status around the forty student threshold until several months into the school year, implying that the LATE may underestimate the average treatment effect away from the threshold. Section 6 discusses the effect of this underestimation on the structural estimates. Equations 4.1 and 4.2 represent sharp RD designs as there is perfect adherence to the subgroup rule under NCLB (see Section 5).

4.2 Identifying Dynamic Complementarities

Single period random assignment does not separately identify each period's treatment effect (or their interactions) since they are perfectly collinear, implying that dynamic complementarities cannot be identified without the year-to-year treatment variation provided by multi-period randomization (Ding and Lehrer, 2010; Almond and Mazumder, 2013). In this subsection, I extend the RD framework, which provides treatment variation that is "as good as randomly assigned" (Lee and Lemieux, 2010), to incorporate the year-to-year treatment variation embedded in the NCLB accountability setting. This yields the multi-period randomization necessary to test whether the student learning technology features dynamic complementarities. Intuitively, the technology features dynamic complementarities when RD estimates are larger among those receiving treatment in the prior period, relative to those who did not.

Define the RD estimand $\tau_t^{(K,...,K,T)}$, where the subscript t denotes the period (where grade 3 is normalized to period $t=1$) and the superscript $(K, ..., K, T)$ denotes the treatment effect in period t for a student with treatment history $(K, ..., K)$, where $K \in \{T, U\}$, with T and U representing the 'treated' and 'untreated' cases, respectively. For exposition, suppose that

the production technology follows the form:[24]

$$f_t(\alpha_i, S_{i1}^K, ..., S_{it}^{(K,...,K)}) = \alpha_i + \sum_{t=1}^{t} S_{it}^{(K,...,K)} + \sum_{j=1}^{t} \beta_{t-j,t} S_{i,t-j}^{(K,...,K)} S_{it}^{(K,...,K)}, \qquad (4.3)$$

where α_i denotes student ability and S_{it} is school inputs at time t with its superscript $(K, ..., K)$ (with $K \in \{T, U\}$) indicating the school inputs for student i in period t with a treatment history of $(K, ..., K)$. This superscript allows for the schooling inputs to differ among students with different treatment histories, which is likely to occur if educators believe that the marginal productivity of investment is higher when students are treated in prior periods. With the above technology, I consider the structural parameters identified by the RD estimands in a two-period framework.

Imagine N students entering school s in period one. Student achievement in period one, $A_{i1}^{(K)}$, can take two possible values: the student can be 'treated' (i.e., held accountable), $A_{i1}^{(T)}$, or 'untreated' (i.e., not held accountable), $A_{i1}^{(U)}$. Given the production technology in Equation 4.3, the period one RD estimand, $\tau_1^{(T)}$, identifies:

$$\tau_1^{(T)} = \frac{1}{N} \sum_{i}^{N} [A_{i1}^{(T)} - A_{i1}^{(U)}] = \frac{1}{N} \sum_{i}^{N} [f_1(\alpha_i, S_{i1}^{(T)}) - f_1(\alpha_i, S_{i1}^{(U)})] = \bar{S_1}^{(T)} - \bar{S_1}^{(U)}, \qquad (4.4)$$

where $\bar{S_1}^{(T)} \equiv \frac{1}{N} \sum_{i}^{N} S_{i1}^{(T)}$ and $\bar{S_1}^{(U)} \equiv \frac{1}{N} \sum_{i}^{N} S_{i1}^{(U)}$ denote average school inputs for 'treated' and 'untreated' students, respectively. Since there are no prior schooling inputs that interact with period one treatment, $\tau_1^{(T)}$ captures the effect of the extra school resources that an accountable student receives, relative to a student that is not held accountable.

In period two (denoted by the '2' subscript), the students from period one advance a period and their (average) achievement takes four possible values: treated in both periods, $\bar{A}_2^{(T,T)}$, untreated in both periods, $\bar{A}_2^{(U,U)}$, treated in period one but not two, $\bar{A}_2^{(T,U)}$, and

[24] Since the regressors can include higher order terms and interaction terms to capture nonlinear relationships, the linearization of the technology imposes few restrictions other than additive separability of ability. In this section, the result follows from the more general technology where possible complementarities among ability and school inputs may exist.

treated in period two but not one, $\bar{A}_2^{(U,T)}$. A period-two RD estimand compares average achievement for students who are treated in period two with those who are not, regardless of prior treatment (i.e. $\bar{A}_2^{(U,T)} + \bar{A}_2^{(T,T)} - [\bar{A}_2^{(U,U)} + \bar{A}_2^{(T,U)}]$). To identify accountability-induced input changes in period two, I separate out the period-two RD estimand into two cases: the student was treated in period one, $\tau_2^{(T,T)}$, and the student was not treated in period one, $\tau_2^{(U,T)}$.

In the joint treatment case, $\tau_2^{(T,T)}$ compares average student achievement when the student is treated in both periods, $\bar{A}_2^{(T,T)}$, to average student achievement when the student is only treated in period one, $\bar{A}_2^{(T,U)}$. Conversely, $\tau_2^{(U,T)}$ compares average student achievement when the student is treated in period two but not one, $\bar{A}_2^{(U,T)}$, to when the student is never treated, $\bar{A}_2^{(U,U)}$. In that case, the estimands identify:

$$\begin{aligned}\tau_2^{(T,T)} &= \bar{A}_2^{(T,T)} - \bar{A}_2^{(T,U)} = f_2(\bar{\alpha}, \bar{S}_1^{(T)}, \bar{S}_2^{(T,T)}) - f_2(\bar{\alpha}, \bar{S}_1^{(T)}, \bar{S}_2^{(T,U)}) \\ &= \bar{S}_2^{(T,T)} - \bar{S}_2^{(T,U)} + \frac{1}{N}\sum_i^N \beta_{12} S_{i1}^{(T)} (S_{i2}^{(T,T)} - S_{i2}^{(T,U)}) \qquad (4.5)\end{aligned}$$

$$\begin{aligned}\tau_2^{(U,T)} &= \bar{A}_2^{(U,T)} - \bar{A}_2^{(U,U)} = f_2(\bar{\alpha}, \bar{S}_1^{(U)}, \bar{S}_2^{(U,T)}) - f_2(\bar{\alpha}, \bar{S}_1^{(U)}, \bar{S}_2^{(U,U)}) \\ &= \bar{S}_2^{(U,T)} - \bar{S}_2^{(U,U)} + \frac{1}{N}\sum_i^N \beta_{12} S_{i1}^{(U)} (S_{i2}^{(U,T)} - S_{i2}^{(U,U)}) \,. \qquad (4.6)\end{aligned}$$

Comparing these two estimands supplies a test for whether the student learning technology features dynamic complementarities:

Proposition 3 $\tau_2^{(T,T)} > \tau_2^{(U,T)}$ *iff* $\beta_{12} > 0$.[25]

Proof. Follows almost immediately from: (i) achievement is increasing in schooling inputs, making $\bar{S}_t^{(T)} > \bar{S}_t^{(U)}$, and (ii) schools respond to the increased marginal benefits under dynamic complementarities by setting $S_{it}^{(T,T)} \geq S_{it}^{(U,T)}$. See Appendix A for the formal proof. ■

It is instructive to consider two cases where $\tau_2^{(T,T)} - \tau_2^{(U,T)} > 0$: (i) when educators do not alter schooling inputs in response to prior treatment histories, and (ii) when educators

[25] A similar proof shows that $\tau_2^{(T,T)} < \tau_2^{(U,T)}$ iff $\beta_{12} < 0$ (i.e., school inputs are dynamic substitutes).

do. When educators do not respond to treatment history – perhaps because they do not know the underlying student learning technology – then it must be that $S_{i2}^{(T,T)} = S_{i2}^{(U,T)}$ and $S_{i2}^{(T,U)} = S_{i2}^{(U,U)}$ $\forall i$. The difference between the two estimators is given by:

$$\tau_2^{(T,T)} - \tau_2^{(U,T)} = \frac{1}{N}\sum_i^N \beta_{12}(S_{i1}^{(T)} - S_{i1}^{(U)})(S_{i2}^{(K,T)} - S_{i2}^{(K,U)}), \text{ where } K \in \{T, U\}\,. \tag{4.7}$$

The difference between $\tau_2^{(T,T)}$ and $\tau_2^{(U,T)}$ therefore simply captures the average interaction between the (student-specific) treatment effects in periods one and two.

When educators respond to prior treatment histories and $\beta_{12} > 0$, they will set $S_{it}^{(T,T)} > S_{it}^{(U,T)}$ since dynamic complementarities make the marginal benefit of period two investment larger when the student is treated in period one. In this case, the difference between $\tau_2^{(T,T)}$ and $\tau_2^{(U,T)}$ is given by:

$$\tau_2^{(T,T)} - \tau_2^{(U,T)} = \bar{S}_2^{(T,T)} - \bar{S}_2^{(U,T)} + \frac{1}{N}\sum_i^N \beta_{12}(S_{i1}^{(T)} - S_{i1}^{(U)})(S_{i2}^{(T,T)} - S_{i2}^{(U,T)})\,, \tag{4.8}$$

since, because period two is terminal, we have that $S_{i2}^{(T,U)} = S_{i2}^{(U,U)} = 0$. The difference between the two estimators is much larger than in Equation 4.7 since this difference incorporates both the direct effect of the dynamic complementarity and school responses to the higher marginal productivity of investment. I use the structural framework in Section 6 to separate out these two components and identify β_{12}.

Identification: Identification of $\tau_1^{(T)}$ is straightforward: since it only requires randomization in period one, the RD regression given by Equation 4.1 – restricted to period one – estimates $\tau_1^{(T)}$. Estimation of $\tau_2^{(T,T)}$ and $\tau_2^{(U,T)}$, however, require randomization in periods one and two. Fortunately, NCLB's year-to-year treatment variation supplies this dual-

randomization, which I implement through the following multi-dimensional RD regression:[26]

$$
\begin{aligned}
y_{sgt} = \gamma_1 + \tau_2^{(U,T)} T_{sgt} + \theta_1 \nu(X_{sgt}, X_{sg,t-1}) + \phi_1(\nu(X_{sgt}, X_{sg,t-1}) * T_{sgt}) + \lambda_g + \theta_t + \delta_1 Z_{sgt} \\
+ D_{sg,t-1}[\gamma_2 + \left[\tau_2^{(T,T)} - \tau_2^{(U,T)}\right] T_{sgt} + \theta_2 \nu(X_{sgt}, X_{sg,t-1}) + \phi_2(\nu(X_{sgt}, X_{sg,t-1}) * T_{sgt}) + \lambda_g \\
+ \theta_t + \delta_2 Z_{sgt}] + \epsilon_{gst}, \quad for \ \ (-w_{sgt}, -w_{sg,t-1}) \leq (X_{sgt}, X_{sg,t-1}) \leq (w_{sgt}, w_{sg,t-1}), \qquad (4.9)
\end{aligned}
$$

where T_{sgt} is an indicator variable equal to one if subgroup g in school s was held accountable in period two, $D_{sg,t-1}$ is an indicator variable equal to one if subgroup g in school s was held accountable in period one, and $\nu(X_{sgt}, X_{sg,t-1})$ is flexible control over the multi-dimensional space defined by the running variables X_{sgt} and $X_{sg,t-1}$, which determine treatment in periods two and one, respectively. All other variables are defined as in Equation 4.2.

The multi-dimensional regression defined by Equation 4.9 uses random treatment variation from periods one *and* two. Identification of $\tau_2^{(U,T)}$ comes from the comparison between two types of school: one with subgroup enrollment of thirty-nine students in period one and forty students in period two, to one with subgroup enrollment of thirty-nine students in periods one and two. Similarly, $\tau_2^{(T,T)}$ is identified by comparing schools with subgroup enrollment of forty in periods one and two to schools with subgroup enrollment of forty in period one and thirty-nine in period two. Thus, the parameters are identified from a student entering or exiting a school randomly in *both* periods one and two. Controlling for subgroup enrollment in *both* periods one and two and restricting the sample to observations with subgroup enrollment within the bandwidth of five in *both* periods ensures that the parameters are identified from this variation.

The above intuition is easily extended to $t = 3$ (i.e. grade 5), giving four estimated moments: $\tau_3^{(U,U,T)}$, $\tau_3^{(T,U,T)}$, $\tau_3^{(U,T,T)}$, and $\tau_3^{(T,T,T)}$. These estimates are estimated through a three-dimensional variant of Equation 4.9. In the three-dimensional design, however, standard errors become very large, making meaningful inference difficult. Therefore, I generally

[26] See Dell (2010), Papay et al. (2014), and Porter et al. (2017) for the nascent literature on identification and estimation in a multi-dimensional RD framework.

neglect the period one treatment status and present these estimates as two-dimensional estimates, $\tau_3^{(U,T)}$, $\tau_3^{(T,T)}$. Finally, following Lee and Card (2008), I cluster standard errors for all RD estimates at the student-by-subgroup level (i.e. by each value of the (discrete) running variable for each subgroup).

Alternative Story: It is possible that $\tau_2^{(T,T)} > \tau_2^{(U,T)}$ if there are dynamic complementarities in the teaching technology, rather than dynamic complementarities in the production technology (and the associated school responses). This alternative story is that treatment in the prior period gives teachers additional skills that make investments more productive in the next period.

I rule out this alternative story by testing for the presence of dynamic complementarities in the teaching technology directly. The test stems from the fact that dynamic complementarities in the teaching technology affect every grade, while dynamic complementarities in the learning technology only affect grades where students can be treated in the prior period. Grade 3 students, who cannot receive treatment in the prior period (but whose teachers taught their treated schoolmates last year), should therefore not exhibit differential treatment effects by prior year treatment status *unless* there is a dynamic complementarity in the teaching technology. Evaluating whether $\tau_1^{(T,T)} > \tau_1^{(U,T)}$ therefore provides a test of the existence of dynamic complementarities in the teaching technology.

4.3 Data

I use detailed administrative data from the North Carolina Education Research Center (NCERDC). These include information about all public school students and teachers in North Carolina for the 2002-03 to 2007-08 school years.[27] Given that NCLB was enacted for the 2002-03 school year, the data cover the first six years of NCLB. They contain test scores for each student in mathematics and reading for grades two through five from standard-

[27] I stop at the 2007-08 school year as a 2008-09 reform allowed students to take two tests to determine whether they were proficient under NCLB, potentially changing the type of students that schools target.

ized tests that are administered at the end of each school year in the state.[28] Test scores are reported on a developmental scale, which is designed such that each additional point represents the same knowledge gain, regardless of the student's grade or baseline ability. To create comparability of test scores across grades, I standardize this scale at the student level to have a mean of zero and a variance of one for each grade-year. Except for Equation 5.1 (non-parametric identification of accountability by ability), all data are collapsed to the 'school-subgroup-belong-year' level, which is the unit of observation in the econometric framework (represented by the 'sbgt' subscript), where each school-subgroup-year observation is split into two separate observations: students who belong and students who do not belong to the subgroup that is being analyzed. As a consequence, these standardized scores no longer have a standard deviation of one.

I obtain data on school-level subgroup counts and the subgroups that are used for NCLB purposes by referring to North Carolina's Adequate Yearly Progress (AYP) reports.[29] The AYP reports include detailed data stating which subgroup-specific targets the school must meet to pass AYP and whether or not the target was actually achieved. The reports also record the number of students in each subgroup who are used for NCLB purposes, allowing me to utilize the understudied forty student subgroup rule for identification. One cannot do this with the NCERDC data set since it reports only the number of students taking the end-of-grade test. The NCLB attendance rules described in Section 2 often make the AYP subgroup counts differ from the number of students belonging to a subgroup who take the end-of-grade test in a given school. The AYP reports are therefore required to obtain consistent school-level subgroup counts for NCLB purposes, a pre-requisite for employing a regression discontinuity design.

The NCERDC data set contains unique student and teacher identifiers, allowing students and teachers to be linked and tracked over time. Classroom assignment data are inferred

[28]Grade 2 tests, which are administered at the start of the grade three school year, are the exception. In addition, grade two math scores for the 2004-05 school year are missing.

[29]As part of NCLB, each state must file these reports. North Carolina's reports are available at `http://accrpt.ncpublicschools.org/app/2003/ayp/`.

based on the teacher who administers the test at the end of each school year. In elementary schools, this is almost always the student's classroom teacher.[30] I make two data restrictions. First, observations are omitted if the NCERDC school-subgroup-belong-year count is less than fifty percent of the count from the AYP data. Second, data are restricted to schools with a highest grade of either 5 or 6 to eliminate K-12 schools, which have especially long time horizons, making it difficult to identify differential investment across grades.[31]

Summary statistics are reported in Table 1. Column (1) shows student characteristics for all students in the sample. North Carolina has a white student plurality and a substantial black minority population (26 percent), with Hispanic and Asian students making up a further fourteen and seven percent of the student body, respectively. Column (2) restricts the sample to observations near the forty-student threshold, which looks very similar to column (1) aside from some minor differences. Columns (3) and (4) report the subsample of students who belong and do not belong to the subgroup that is near the forty-student threshold, respectively. There are more substantial differences here, with students belonging to the subgroup scoring about 0.3σ lower on the end-of-grade standardized tests than students not belonging to the subgroup. Correspondingly, students belonging to the subgroup are much more likely to be black, Hispanic or disadvantaged. These differences are an artifact of the sampling: schools are far more likely to have a minority subgroup close to the forty student subgroup threshold than a majority subgroup. Appendix Table A3 also reports differences in test scores across subgroups, showing that black and disadvantaged students perform far worse on standardized tests than white or Asian students (as found elsewhere).

5 Results

I discuss the validity of the regression discontinuity design introduced in Section 4 and present the reduced-form results. Then, building on the period-by-period treatment variation

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

[31]These restrictions eliminate about one and three percent of schools in the RD sample, respectively.

(as described in subsection 4.2), I formally test for dynamic complementarities among school inputs. Section 6 then utilizes these results in the context of a structural model.

5.1 Validity of the Regression Discontinuity Design

Because schools do not face any subgroup-specific accountability if there are fewer than forty students in a subgroup, they have an incentive to manipulate the number of students in subgroups. As schools that manipulate their subgroup numbers may systematically differ from those that do not, such manipulation may invalidate the RD design.

Relevant to the manipulation concern, Section 2 describes how the number of students in a given subgroup is calculated. Schools may be able to manipulate the number of students in a subgroup in a variety of ways: changing a student's subgroup designation, refusing admission to a student of a certain subgroup, or preventing a student from reaching 140 days at the school through suspension or expulsion. For racial subgroup categories, changing a student's racial designation seems difficult, but there is some evidence that schools are able to change a student's disability designation;[32] in light of that, the disability subgroup is omitted from the analysis that follows.

To check whether schools manipulate the number of students in a subgroup, Figure 2 plots the distribution of the number of students by school-subgroup-year cell. If schools are manipulating the number of students in a subgroup, 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 the visual analysis: the null hypothesis of continuity at the cutoff is not rejected.[33]

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 I control

[32]See Jacob (2005). There is no evidence in North Carolina of manipulation of the other two non-racial subgroups of limited English proficient or disadvantaged.

[33]Similarly, I conduct a McCrary (2008) test for the four largest subgroups, separately. This yields p-values of 0.92, 0.72, 0.37 and 0.21 for the black, Hispanic, white and disadvantaged subgroups, respectively.

for observable characteristics, discontinuities in observable characteristics may be suggestive of changes in unobservables around the cutoff. Figures 3 and 4 graph mean covariates by the number of students in a subgroup for students belonging to and not belonging to a subgroup, respectively. These covariates include grade 2 math and reading scores, race, free lunch status, limited English-proficient status, gifted status and disability status. The covariates appear smooth around the cutoff. More formally, Appendix Table A1 estimates the discontinuity in covariates at the threshold. Panel A reports the results for students belonging to the subgroup, Panel B for students not belonging to the subgroup, and Panel C reports the difference in these, as the difference-in-discontinuity design (see Equation 4.2) requires no discontinuities in the differences of unobservables characteristics. As expected, the covariates do not exhibit a significant discontinuity at the cutoff (in line with expectations), though a few covariates are significant at the ten percent level (four out of thirty-three). Finally, I implement a seemingly unrelated regression for each panel to test the null hypothesis that all covariates are jointly continuous at the threshold. Given the p-values, the null hypothesis in each panel cannot be rejected.

The validity of the two-dimensional RD design requires no discontinuities in observable or unobservable characteristics in both dimensions. Appendix Table A2 demonstrates that there are no observable discontinuities in covariates at the threshold for period two students who were and who were not treated in period one (or a difference-in-discontinuity between them).[34] In addition, to ensure that heterogeneous treatment effects are not being attributed to dynamic complementarities, Panel D of Appendix Table A2 shows that covariate levels are similar between period two students who were and who were not treated in the prior year.

For the regression discontinuity design to be valid, having more than forty students in a subgroup must cause schools to be held accountable under NCLB for that subgroup. Figure 5 plots whether or not the subgroup-specific accountability target was used for NCLB purposes

[34] Appendix Table A2 does this only for students in period two as this time period generates the key results in this paper. A similar check for students in period three is available upon request.

by the number of students in that subgroup. As expected, once the forty student threshold is crossed, schools are held accountable for that subgroup's performance. In fact, adherence to the rule is perfect, with a one-hundred percent jump in the probability that a school is held accountable for that subgroup once the threshold is crossed.[35]

5.2 Results from the RD Design

Figure 6 plots the reduced-form relationship between the number of students in a subgroup and student achievement both for students who belong and do not belong to the subgroup. As expected, students belonging to the subgroup see their math scores jump when the forty student threshold is crossed. This jump is not observed for students who do not belong to the subgroup. The reduced-form relationship for reading scores is less clear: a similar large jump is observed both for students belonging and not belonging to the subgroup. The discontinuity shrinks, however, when additional controls are added, consistent with the NCLB literature, which finds smaller responses for English test scores.[36] The point estimates above the subfigures correspond to the estimates of $\tau_{i \in g}$ and $\tau_{i \notin g}$ in columns (1) and (3) of Table 2.

Since schools only pass AYP if every subgroup attains its proficiency standard, certain subgroups, especially the lowest performing subgroup, are more pertinent as to whether a school passes AYP. Appendix Table A3 shows that, on average, the black and disadvantaged subgroup are low-performing, while the white subgroup is higher-performing. In light of this evidence, school responses should be larger for the black or disadvantaged subgroup relative to the white subgroup. Appendix Figures A1 and A2 plot the reduced-form relationships between the number of students in the black, disadvantaged and white subgroups – together

[35]The jump in probability is not observed for limited English-proficient students. This is likely because 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. In fact, there is no observable discontinuity around the forty student threshold for this subgroup and it is therefore not used in this paper.

[36]For example, Dee and Jacob (2011) conclude that NCLB increases math scores but not English scores. There could be two reasons for this: either school have difficulties influencing English scores, or math targets are more binding than English ones. Although both reasons may be operating, there is clear support for the latter reason in North Carolina: twice as many schools fail AYP in math than in English.

representing about seventy percent of the RD sample – and mean standardized test scores for math and reading, respectively. As expected, there is a jump in test scores for students belonging to the black and disadvantaged subgroups once the forty student threshold is crossed, while no such jump occurs for students who do not belong to the black or disadvantaged subgroups. The white subgroup, on the other hand, has no obvious discontinuity in outcomes once the forty student threshold is crossed.

Table 2 reports the results from Figure 6 in columns (1) and (3) – with controls being added in columns (2) and (4) – for math and reading, respectively. Column (2) of Panel A shows that students belonging to an accountable subgroup have significantly higher math achievement. With controls, the estimated effect of subgroup accountability on students in that subgroup is 0.053σ, which is significant at the five percent level. Panel B reports the effect for students not belonging to the subgroup, which is close to zero and not statistically significant. Taking the difference in these outcomes (Panel C), we conclude that subgroup-specific accountability increases math achievement for students belonging to the accountable subgroup by 0.046σ relative to students not belonging to that subgroup, significant at the one percent level. Panel A of column (4) shows that subgroup-specific accountability increases reading scores among students belonging to that subgroup by 0.028σ, which is about half the observed effect on math scores. For students not belonging to the subgroup (Panel B), the effect of subgroup accountability on reading scores is not statistically significant. Henceforth, I focus on math scores to leverage their larger effects in the counterfactual analysis. In addition, I do not consider spillovers in the counterfactual analysis, consistent with the small and statistically insignificant effects on students not belonging to the subgroup.[37]

5.3 Dynamic Complementarities

From subsection 4.2, recall that $\tau_t^{(T,T)}$ and $\tau_t^{(U,T)}$ represent the treatment effect among students that were 'treated' and 'untreated' in the prior period, respectively. One might

[37]Since the estimates indicate that spillovers are positive, the inclusion of spillovers would increase the benefit of any higher performing incentive scheme in the counterfactual analysis.

expect that the average of these estimates should approximately equal the estimates in Table 2: however, the two-dimensional RD restriction requires schools to be near the forty student threshold in both the current and prior period. This implies that a different LATE is identified here: schools consistently near the forty student threshold period-by-period may respond differently than schools that are not. Specifically, schools that are not consistently near the threshold are likely experiencing large changes to their subgroup enrollment: these schools may respond less to crossing the forty student threshold as in the future they expect to be always (never) held accountable as their subgroup enrollment is on a upward (downward) trajectory, consistent with evidence from Subsection 8.1 that schools form expectations over future treatment.

In addition, since schools have the same student over multiple periods, a theoretical model with a cumulative production technology (see Section 3) would indicate that schools would invest earlier in students since this raises the probability they exceed the fixed NCLB threshold both today *and* in future periods. Consequently, we expect estimates to be larger in earlier grades than in later grades. This is consistent with the period-specific RD estimates provided in Appendix Table A5.

Table 3 provides period-specific estimates of $\tau_t^{(T,T)}$ and $\tau_t^{(U,T)}$ using the two-dimensional RD design described in Equation 4.9. I start by focusing on the period one results, which act as a 'placebo' test by checking for dynamic complementarities in the *teaching* technology (since there was treatment for these teachers in the prior year, but not students), and then turn to periods two and three which provide a direct test of dynamic complementarities in the *learning* technology.

Testing for Dynamic Complementarities in the *Teaching* Technology: Students that are entering period one (i.e., grade 3) are not subject to the accountability scheme in the prior year, implying that the school's prior year treatment status cannot generate dynamic complementarities in the learning technology. Panel A of Table 3 shows that the difference between $\hat{\tau}_1^{(T,T)}$ and $\hat{\tau}_1^{(U,T)}$ is close to zero ($\hat{\tau}_1^{(T,T)} - \hat{\tau}_1^{(U,T)} = 0.028$ [$s.e.$ 0.119]); the treatment

effect for period one therefore does not depend on prior treatment status. Since these teachers would have taught a 'treated' class the prior year, if the dynamic complementarity existed through the teaching technology (rather than the learning technology) then we would expect $\hat{\tau}_1^{(T,T)} > \hat{\tau}_1^{(U,T)}$ (as we will see for periods two and three).

Testing for Dynamic Complementarities in the *Learning* Technology: Since students who were treated in the prior period received more inputs relative to those who were untreated, $\tau_t^{(T,T)} > \tau_t^{(U,T)}$ indicates the presence of dynamic complementarities (Proposition 3). In Table 3, I therefore test whether $\hat{\tau}_2^{(T,T)} > \hat{\tau}_2^{(U,T)}$ and $\hat{\tau}_3^{(T,T)} > \hat{\tau}_3^{(U,T)}$.[38] In period two, the difference between these estimates is 0.189σ and is statistically significant at the ten percent level, while in period three the difference is 0.172σ (p-value=0.216). Panel D of Table 3 combines these results by taking their arithmetic average. The estimate here is 0.180σ and is significant the five percent level (p-value=0.044), providing evidence for a dynamic complementarity between school inputs in the learning technology.

5.4 Testing Model Predictions

I test the two theoretical predictions of the model by adapting the regression discontinuity framework introduced in Section 4.[39]

Proposition 1: Without dynamic complementarities, one could test Proposition 1 by estimating Equation 4.1 separately for each grade (as done in Appendix Table A5): under Proposition 1, the effect of subgroup-accountability should be higher in earlier grades. If there are dynamic complementarities, however, these grade-specific estimates do not capture the effect of accountability-induced increases to school inputs in each grade. The intuition is that when students are treated for t periods, the period t RD estimator captures higher levels

[38] Recall that I use a two-dimensional rather than a three-dimensional RD design in the three period case due to a limited sample size: the three-dimensional RD design has only 125 observations. Estimates (s.e.'s) for the three-dimensional design are: $\hat{\tau}_3^{(U,U,T)}$=−0.317 (0.371), $\hat{\tau}_3^{(T,U,T)}$=−0.419 (0.437), $\hat{\tau}_3^{(U,T,T)}$=0.363 (0.404), and $\hat{\tau}_3^{(T,T,T)}$=0.260 (0.492).

[39] I defer testing the corollary of Proposition 1 until Section 8, since it is better placed in a wider discussion of structural assumptions.

of period t school inputs *and* their dynamic interactions with the higher levels of prior inputs for these students. Therefore, period-specific RD estimates (except period one) identify a combination of period-specific input changes and dynamic complementarities.

I test Proposition 1 by checking whether $\tau_1^{(T)} > \tau_2^{(U,T)} > \tau_3^{(U,T)}$. This test comes from the fact that the dynamic complementarity term in $\tau_2^{(U,T)}$ (given by Equation 4.5) is an interaction between period one inputs for an 'untreated' student and her period two treatment effect. While period one school inputs for untreated students, $S_1^{(U)}$, are non-zero since the school foresees that the student may be treated in period two, they are likely to be small (this holds true in the structural analysis that follows), making the dynamic complementarity term in $\tau_2^{(U,T)}$ a small positive value. Therefore, $\tau_1^{(T)} > \tau_2^{(U,T)}$ is a necessary (but not sufficient) condition for Proposition 1 that schools invest more in period one than in period two (i.e., $S_1^{(T)} > S_2^{(U,T)}$). A similar argument confirms that $\tau_2^{(U,T)} > \tau_3^{(U,T)}$ tests whether schools invest more in period two than in period three.

Table 3 provides estimates of $\tau_1^{(U,T)}$, $\tau_2^{(U,T)}$, and $\tau_3^{(U,T)}$. In line with Proposition 1, the estimates indicate that schools invest more in earlier periods since $\hat{\tau}_1^{(U,T)} > \hat{\tau}_2^{(U,T)} > \hat{\tau}_3^{(U,T)}$ (recall that this is a necessary condition for Proposition 1 from Subsection 5.4). While the difference between $\hat{\tau}_1^{(U,T)}$ and $\hat{\tau}_2^{(U,T)}$ is not significant, tests for $\hat{\tau}_1^{(U,T)} > \hat{\tau}_3^{(U,T)}$ (p-value=0.013) and $\hat{\tau}_2^{(U,T)} > \hat{\tau}_3^{(U,T)}$ (p-value=0.012) show that investments in period one and period two are significantly higher than in period three, providing empirical support for Proposition 1.

Proposition 2: Proposition 2 establishes that schools target students with ability levels below the achievement threshold under a fixed-target scheme. I investigate this by first predicting student ability, α_i, with pre-accountability test scores (i.e., test scores in grade 2) and a full set of demographic controls,[40] yielding a predicted ability level, $\hat{\alpha}_i$. I then estimate the treatment effects for each ability level. I identify these ability-specific estimates, $\tau_1^{(T)}(\alpha_i)$ and $\tau_2^{(T,T)}(\alpha_i)$, through nonparametric RD regressions that vary by a student's predicted

[40]Formally, I regress $\alpha_i = \psi_0 + \psi_1 y_{i,g=2} + \psi_3 Z_i + \epsilon_i$, where $y_{i,g=2}$ is grade 2 test scores and Z_i are the demographic characteristics of student i.

ability level. For example, I estimate $\tau_1^{(T)}(\alpha_i)$ with:

$$y_{isgt}(\alpha_i) = \gamma_0 + \tau_1^{(T)}(\alpha_i)T_{sgt} + \theta X_{sgt} + \phi(X_{sgt} * T_{sgt}) + \lambda_g + \theta_t + \delta Z_{isgt} + \epsilon_{igst}\,,$$
$$\forall gst \in [-w_{sgt}, w_{sgt}], \quad \forall i \in [\alpha_i - \kappa_i, \alpha_i + \kappa_i]\,, \tag{5.1}$$

where w_{sgt} is the usual RD bandwidth, κ_i is the bandwidth with respect to student ability (set to one standard deviation), and all other variables are defined in Equation 4.1. Regressions are weighted by ability through a triangular kernel that puts more weight on observations closer to $y_{isgt}(\alpha_i)$. Under Proposition 2, there should be an 'inverse-U' shaped distribution of ability-specific treatment effects that is centered below the achievement threshold.

Figure 7 displays how the treatment effects vary by a student's predicted ability. Figures 7(a) and 7(b) report the results from Equation 5.1, which estimates the periods one and two RD estimates, $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$, for various predicted student ability levels, $\hat{\alpha}_i$, respectively. Consistent with Proposition 2, the figures suggest that schools target students who are predicted to be below-marginal with respect to NCLB's achievement threshold. Furthermore, the fact that the 'inverted U' is located similarly in Figures 7(a) and 7(b) indicate that schools re-invest in the same below-marginal students in period two,[41] suggesting that educators utilize dynamic complementarities to maximize period two student achievement and thus have a sense of the underlying technology.

6 Structural Estimation

The theoretical model in Section 3 explored school responses to a fixed-target scheme from a dynamic perspective. Now, I turn to estimating the key model parameters, $\beta_{t-1,t}$ and $\frac{\psi}{b}$, which govern dynamic complementarities and the cost-to-benefit ratio of school inputs,

[41] Recall that predicted ability incorporates period 0 (i.e., grade 2) test scores and demographic characteristics only for both the period one and two figures. These figures are therefore not at odds with those in Neal and Schanzenbach (2010), who look at period two test scores as a function of ability predicted using period one test scores.

respectively. To do so, I follow Aizer and Cunha (2012) and assume that the technology in Equation 3.1 takes the following functional form:

$$A_{it} = \alpha_i + S_{it} + S_{i,t-1} + \beta_{t,t-1} S_{it} S_{i,t-1} + \epsilon_{it} \,. \tag{6.1}$$

Because of power issues, I consider a two-period model (i.e. $T = 2$) and set grades 3 and 4 to correspond to time periods one and two, respectively.[42] The human capital technology is therefore given by:

$$A_{i1} = \alpha_i + S_{i1}^{(K)} + \epsilon_{i1} \,, \tag{6.2}$$

and

$$A_{i2} = \alpha_i + S_{i1}^{(K)} + S_{i2}^{(K,K)} + \beta_{12} S_{i1}^{(K)} S_{i2}^{(K,K)} + \epsilon_{i2} \,, \tag{6.3}$$

allowing schooling inputs to depend on students' treatment statuses in periods one and two with the superscript (K) and (K, K) (with $K \in \{T, U\}$).

Recall from Section 3 that $\Gamma_{i2} = 1$ denotes the case where student i is treated in period two and $\Gamma_{i2} = 0$, the case where student i is untreated in that period, where $\mathbb{E}[\Gamma_{i2} = 1] = p_i$ under rational expectations. Since the uncertainty is realized in period two, schools make period two input decisions given the realization of Γ_{i2} and their period one decisions. In period one, schools determine inputs given their expectations over which period two outcome will be realized. Since the school decides sequentially, the model is solved by backward induction.

Period Two: The school's problem depends on the treatment realization, Γ_{i2}, and period one inputs. When $\Gamma_{i2} = 1$ (i.e., the school is held accountable), the school solves:

$$\begin{aligned} \underset{\{S_{i2}^{(K,T)}\}_{K \in \{T,U\}}}{max} & \sum_{i}^{N} b\Big[H(\alpha_i + S_{i1}^{(K)} + S_{i2}^{(K,T)} + \beta_{12} S_{i1}^{(K)} S_{i2}^{(K,T)} - \mathcal{A}_2^*) - c(S_{i1}^{(K)}, S_{i2}^{(K,T)})\Big] \\ & \textit{subject to: } S_{i2}^{(T,T)} \geq 0, \; S_{i2}^{(U,T)} \geq 0 \;\; \forall i \,. \end{aligned} \tag{6.4}$$

[42] Estimation of a three-period model requires estimates of $\tau_3^{(T,T,T)}$, $\tau_3^{(T,U,T)}$, $\tau_3^{(U,T,T)}$ and $\tau_3^{(U,U,T)}$. These four reduced-form estimates are found in footnote 38; however, they are estimated using a small sample since they condition on treatment statuses two years prior, eliminating a year of data and requiring schools to be near the threshold for three consecutive years (estimates rely on 125 school-subgroup-year observations). I therefore lack the precision in the current application to identify a three-period model.

If the student is not held accountable (i.e., $\Gamma_{i2} = 0$), then period two inputs are trivially set to zero, since there are no future benefits to raising school inputs in period two (given it is a two-period model).

Period One: Given rational expectations, the school's problem for $K \in \{T, U\}$ is:

$$\underset{S_{i1}^{(K)}}{max} \sum_{i}^{N} \Big[\Gamma_{i1} \cdot b \cdot H(\alpha_i + S_{i1}^{(K)} - \mathcal{A}_1^*) + p_i \cdot b \cdot H(\alpha_i + S_{i1}^{(K)} + S_{i2}^{(K,T)} + \beta_{12} S_{i1}^{(K)} S_{i2}^{(K,T)} - \mathcal{A}_2^*) - c(S_{i1}^{(K)}, S_{i2}^{(K,T)}) \Big]$$

$$\textit{subject to: } S_{i1}^{(K)} \geq 0 \ \ \forall i \,, \tag{6.5}$$

Taking the first-order conditions (FOCs) from the school problem in Equations 6.4 and 6.5 with respect to school inputs in period one and two yields the implied optimal level of school inputs implicitly satisfying the following equations:

$$c'(\cdot) = \begin{cases} b \left[h(\alpha_i + S_{i1}^{(T)} - \mathcal{A}_1^*) + p_i \cdot h(\alpha_i + S_{i1}^{(T)} + S_{i2}^{(T,T)} + \beta_{12} S_{i1}^{(T)} S_{i2}^{(T,T)} - \mathcal{A}_2^*) \cdot (\beta_{12} S_{i2}^{(T,T)}) \right] & \text{if } \Gamma_{i1} = 1 \\ p_i \cdot b \cdot h(\alpha_i + S_{i1}^{(U)} + S_{i2}^{(U,T)} + \beta_{12} S_{i1}^{(U)} S_{i2}^{(U,T)} - \mathcal{A}_2^*) \cdot (\beta_{12} S_{i2}^{(U,T)}) & \text{if } \Gamma_{i1} = 0 \,, \end{cases} \tag{6.6}$$

$$c'(\cdot) = \begin{cases} b \left[h(\alpha_i + S_{i1}^{(K)} + S_{i2}^{(K,T)} + \beta_{12} S_{i1}^{(K)} S_{i2}^{(K,T)} - \mathcal{A}_2^*) \cdot (\beta_{12} S_{i1}^{(K)}) \right] & \text{if } \Gamma_{i2} = 1 \\ 0 & \text{if } \Gamma_{i2} = 0 \,. \end{cases} \tag{6.7}$$

While neither the cost parameter, ψ, nor the benefit parameter, b, is separately identified, I can identify the marginal cost-to-benefit ratio, $\frac{\psi}{b}$. To solve for the parameters β_{12} and $\frac{\psi}{b}$, I: (1) parameterize the cost function as follows: $c(S_{i1}, S_{i2}) = \frac{\psi}{2}(S_{i1}^2 + S_{i2}^2)$, where ψ governs the slope of the marginal cost curve; (2) proxy initial ability, α_i, using predicted test scores, $\hat{\alpha}_i$, based on grade 2 test scores and a full set of student demographic characteristics; (3) assume that the distribution of the error term is drawn from a mean-zero normal distribution with a variance equal to the variance of the test score in each grade, conditional on initial ability;[43] and (4) assume that schools form their beliefs based on the empirical probability that the relevant school subgroup is treated in period two, given their current student population.[44]

[43] The conditional test score has a variance of approximately 0.85 in each grade.

[44] A school's ability to foresee its period two accountability status depends upon the propensity for students to switch schools and year-to-year variation in subgroup size of the entering cohort. Section 8.1 shows that

To identify the parameters, I use two reduced-form estimates from Section 5: the effect of being treated in period one, $\tau_1^{(T)}$, and the effect of being treated in period two, conditional on also being treated in period one, namely $\tau_2^{(T,T)}$. It is clear from Equations 6.4 and 6.5 (and Proposition 2) that school input decisions depend upon student ability. The model thus requires consistent ability-specific estimates, $\tau_1^{(T)}(\alpha_i)$ and $\tau_2^{(T,T)}(\alpha_i)$, which Equation 5.1 identifies through nonparametric RD regressions that vary by a student's predicted ability level.

Consider these treatment effects in terms of period one and two student achievement, given by Equations 6.2 and 6.3. $\tau_1^{(T)}(\alpha_i)$ compares period one achievement for those who are held accountable – thus receiving school inputs $S_{i1}^{(T)}(\alpha_i)$ – and those who are not held accountable (and therefore receiving the minimum level of school inputs). Similarly, $\tau_2^{(T,T)}(\alpha_i)$ captures the effect of period two treatment, conditional on receiving treatment in period one. Thus, these two empirical estimates can be written in terms of the model as follows:

$$\tau_1^{(T)}(\alpha_i) = \tilde{S}_{i1}^{(T)}(\alpha_i) \tag{6.8}$$

$$\tau_2^{(T,T)}(\alpha_i) = \tilde{S}_{i2}^{(T,T)}(\alpha_i) + \beta_{12}\tilde{S}_{i1}^{(T)}(\alpha_i)\tilde{S}_{i2}^{(T,T)}(\alpha_i)\,, \tag{6.9}$$

where $\tilde{S}_{it}^{(T)}(\alpha_i)$ represent the increased level of school inputs when the school is treated, relative to untreated, in period t.[45] These inputs are an explicit function of student ability.

Using the mapping between the reduced-form estimates and school inputs, I solve for the structural parameter β_{12} by taking the ratio of Equations 6.6 and 6.7 and using Equations 6.8 and 6.9 to replace the school inputs, $\tilde{S}_{i1}^{(T)}(\alpha_i)$ and $\tilde{S}_{i2}^{(T,T)}(\alpha_i)$, with the empirical estimates $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ for each $\hat{\alpha}_i$. I then solve the system of two non-linear equations with

schools can use grade-specific enrollment levels to form their beliefs about future treatment. In my setting, however, there remains substantial year-to-year treatment variation: among treated school-subgroups in my RD sample, only sixty percent of them are treated in period two.

[45] When a school is untreated in period two, the school trivially does not invest any inputs in its students (see the untreated case in Equation 6.7). I pin down the level of inputs for the untreated case in period one using $\hat{\tau}_2^{(U,T)}$. Since this estimate is close to zero over the relevant ability range, $S_{i1}^{(U)}$ is approximately zero in the untreated case.

two unknown parameters, $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$.

Identification: Since β_{12} is estimated for each ability level,[46] I set the estimated parameter $\hat{\beta}_{12}$ to the average of the estimated $\hat{\beta}_{12}(\hat{\alpha}_i)$. The parameter $\hat{\beta}_{12}$ is implicitly identified by the ratio of $S_{i1}^{(T)}$ and $S_{i2}^{(T,T)}$. As $S_{i1}^{(T)}$ approaches $S_{i2}^{(T,T)}$, the estimated $\hat{\beta}_{12}$ becomes larger since a school input profile where $S_{i1}^{(T)} \approx S_{i2}^{(T,T)}$ suggests that the school is leveraging the dynamic complementarity term in the period two achievement equation.[47] When the ratio implies that $S_{i1}^{(T)} > S_{i2}^{(T,T)}$ by a large margin, the estimated $\hat{\beta}_{12}$ becomes small since the school is not leveraging any dynamic complementarities.

Intuitively-speaking, the marginal cost-to-benefit ratio, $\frac{\psi}{b}$, identifies the overall level of school investment since a lower cost (or a higher benefit) causes the school to invest more. This parameter is required for the counterfactual analysis that follows and is solved for by substituting $\hat{\beta}_{12}$ into either Equation 6.6 or 6.7 – either equation yields an identical estimate.

Section 4.1 raised the concern that the LATE in the RD design – which identifies $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ – might understate the true reduced-form effect. If this is the case, then both structural parameters are *underestimated*. This arises because when both $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ are underestimated (by an equal proportion), school inputs in period one, $S_{i1}^{(T)}$, fall more than school inputs in period two, $S_{i2}^{(T,T)}$. Since $S_{i1}^{(T)} > S_{i2}^{(T,T)}$ (by Proposition 1), $S_{i1}^{(T)}$ therefore moves closer to $S_{i2}^{(T,T)}$. From the argument above, this implies that the structural model predicts a higher level of dynamic complementarity since the period-specific inputs are closer to being equalized. The parameters are not overly sensitive to this underestimation, however: if both reduced-form estimates are understated by ten percent, then the structural parameters increase by five percent.

[46] Variation in β_{12} with respect to student ability could be used to identify an interaction between student ability and school inputs in the production technology. In practice, identification of this interaction is infeasible in this context, since schools target a relatively narrow ability range under a fixed-target scheme.

[47] This follows because the dynamic complementarity term in the production technology is multiplicative, implying that period two achievement is maximized when $S_{i1}^{(T)} = S_{i2}^{(T,T)}$. The production technology captures the idea that "if investments complement each other strongly, optimality implies that they should be equal in both periods" (Cunha et al., 2010).

6.1 Estimated Structural Parameters

Figure 7 shows the ability-specific estimates of $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ from Equation 5.1. While $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ maps directly into accountability-induced changes to period one school inputs, recall that $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ captures accountability-induced period two inputs and dynamic complementarities between $S_1^{(T)}$ and $S_2^{(T,T)}$. Therefore, it is unsurprising that $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i) > \hat{\tau}_1^{(T)}(\hat{\alpha}_i)$: this suggests that dynamic complementarities are non-negligible.

$\hat{\beta}_{12}$ is estimated for each predicted ability, $\hat{\alpha}_i \in [-1.2, 0]$. As expected, the parameter is large and stable over the ability distribution where schools respond.[48] Taking an average over this range yields an estimate of $\hat{\beta}_{12} = 10.24$ (s.e. 3.22). Substituting the estimated $\hat{\beta}_{12}$ into either Equation 6.6 or 6.7 yields $\frac{\hat{\psi}}{b} = 10.10$ (s.e. 1.55). Standard errors for these parameter estimates are bootstrapped.

Interpretating the structural dynamic complementarity parameter (as for any interaction term) must be placed in context of the distribution of period one and two accountability-induced changes to school inputs. The estimated dynamic complementarity parameter, $\hat{\beta}_{12}$, indicates that a one standard deviation increase in these period one and two inputs (about 0.08σ in total) leads to a 0.015σ increase in student achievement through dynamic complementarities. To put this into perspective, the dynamic complementarity portion of the production technology provides an average achievement increase of 0.04σ under NCLB, and the parameter $\frac{\psi}{b}$ identifies the *level* of school inputs under the NCLB benefit scheme. This latter parameter allows us to perform counterfactual analysis for alternative accountability schemes, where the benefits under these schemes are similar to those under NCLB.[49]

[48] $\hat{\beta}_{12}$ ranges between five and fifteen over the ability range $\hat{\alpha}_i \in [-1.2, 0]$. When $\hat{\alpha}_i$ goes above zero, this is no longer true as there are no school responses in this region to identify $\hat{\beta}_{12}$. This equates to identifying β_{12} by dividing a zero by a zero, making $\hat{\beta}_{12}$ go towards infinity.

[49] The counterfactual analysis requires that benefits under alternative schemes are anchored to the benefits of NCLB (which is not a monetary scheme). Despite that, Macartney et al. (2016) monetize NCLB sanctions, allowing policymakers to anchor benefits to NCLB when constructing alternative schemes.

6.2 Robustness

I provide numerous robustness checks for both the reduced-form estimates and structural parameters from Sections 5 and 6. A bandwidth of five is used throughout this paper. Appendix Figure A3 tests the sensitivity of the reduced-form estimates, $\hat{\tau}_{i \in g}$, and the estimated structural parameters, $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$, with respect to the bandwidth. Specifically, Appendix Figure A3 plots the resulting estimate for each bandwidth $w_{sgt} \in [2, 10]$, showing that the estimates do not vary substantially.

Estimates in the paper are generated with a triangular kernel functional form, allowing for different functions on either side of the cutoff. The relationship between student outcomes and the number of students in a subgroup around the discontinuity may take a different functional form, however. Appendix Table A4 reports reduced-form and parameter estimates for two additional functional forms: linear and quadratic. The different functional forms yield an effect that is qualitatively and quantitatively similar to estimates using the triangular kernel functional form.

The underlying assumption of the regression-discontinuity design is that outcomes would be continuous through the threshold in the absence of the subgroup rule. While indirect tests of this assumption are provided in Section 5.1, no direct test is available. Nevertheless, smoothness of outcomes over placebo subgroup rules are indicative of smoothness at the true threshold in the absence of the subgroup rule. Appendix Figure A4 reports the reduced-form estimates over a range of placebo subgroup rules. In general, the vast majority of subgroup rules yield an estimate near zero and thus outcomes appear smooth, except at the actual subgroup rule of forty students.

7 Counterfactual Analysis

Having estimated the structural parameters, I can now consider the effects of alternative accountability schemes. First, I place the policy alternatives in the context of the theoretical

model in Section 3. Second, I derive optimal school responses as a function of the structural parameters using first-order conditions from the theoretical model. Those two components in hand, the estimated structural parameters and dynamic model allow me to simulate the full distribution of student achievement under alternative schemes. With this dynamic framework, I highlight various counterfactuals, computing student achievement under no accountability scheme, fixed-target schemes with different achievement thresholds, and value-added schemes, which set student-specific targets based on prior test scores. The counterfactual framework emphasizes that value-added schemes setting student-specific achievement targets based on *baseline* test scores – which I call 'multiperiod value-added' schemes – can *both* increase achievement and reduce inequality.

7.1 Counterfactual Setup

Since accountability schemes usually encompass all students within a school or region,[50] I assume that accountability schemes apply to all students across all time periods in the counterfactual world, so that $p_i = 1$, $\forall i$. From Section 3, the school's optimization problem thus becomes:

$$\underset{\{S_{it}\}_{i \in N}}{max} \sum_{i}^{N} \sum_{t}^{T} [b \cdot H(A_{it} - \mathcal{A}^*) - c(S_{it})]$$

$$\textit{subject to:} \qquad S_{it} \geq 0 \qquad \forall i, t\,. \tag{7.1}$$

Moving the Achievement Threshold: If the policymaker wishes to maintain the fixed-target scheme, she can alter school incentives (and thus the distribution of student achievement) by setting an achievement target of $\mathcal{A}^* + \delta$, where δ can either be positive or negative.[51] Since the FOCs take a similar form to the NCLB case, school responses are easily derived using the theoretical model in Section 3. Given an achievement target $\mathcal{A}^* + \delta$,

[50] In the structural analysis, treatment expectations were important since treatment effects were identified among schools where some students would not be held accountable. Even under NCLB, only a small minority (less than ten percent) of the student population does not face subgroup accountability.

[51] In the counterfactuals I do not vary the achievement threshold by grade (i.e. $\mathcal{A}_1^* = \mathcal{A}_2^*$). This is just for expositional clarity; I can simulate the achievement distribution under grade-specific achievement thresholds.

I can therefore use the estimated parameters $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$ to calculate school inputs $S_{i1}(\hat{\alpha}_i)$ and $S_{i2}(\hat{\alpha}_i)$ $\forall i$ based on Equations 6.6 and 6.7. With the school inputs in hand, I then simulate the distribution of student achievement under alternative fixed-target schemes, parametrized by δ.

Value-added Schemes: Value-added schemes, which set student-specific achievement thresholds, generate a substantial change to school incentives. A value-added scheme specifies the achievement level threshold of student i, $\mathcal{A}^*_{it}$, based on test scores from a prior period plus a test score improvement target, ζ_k. The target is now both student- and year-specific and is given by: $\mathcal{A}^*_{it} = A_{i,t-k} + \zeta_k$, where $k \geq 1$, $\zeta_k \geq 0$ and $A_{i,t-k} = A_{i0}$ if $t < k$. The parameter k determines which prior test score the target is based on. The policymaker can set a different test score improvement target, ζ_k, based on the type of value-added scheme denoted by the parameter k.[52] I differentiate between a 'traditional value-added' scheme, which sets the achievement threshold based on last year's test score (i.e. $k = 1$) and 'multiperiod value-added' schemes where $k \geq 2$.[53] As in the fixed-target scheme, schools get benefits b based on the proportion of students who exceed this threshold and they face a convex cost of resources, $c(S_{it})$. The school's optimization problem under a value-added scheme is:

$$\max_{\{S_{it}\}_{i \in N}} \sum_{i}^{N} \sum_{t}^{T} \left[b \cdot H(A_{it} - A_{i,t-k} - \zeta_k) - c(S_{it}) \right]$$

$$\textit{subject to:} \qquad S_{it} \geq 0 \qquad \forall i, t\,, \tag{7.2}$$

[52] Indeed, the policymaker should do so since ζ_k represents the test score improvement target over k periods. If the policymaker wants a test score improvement of 0.1 units over two periods, she must set $\zeta_1 = 0.05$ and $\zeta_2 = 0.1$ for value-added schemes based on lagged and two year-lagged test scores, respectively.

[53] In my context, this involves using the test score at the start of accountability (as there is no accountability or tests before grade three). This is also the entering test score for many schools.

with first-order conditions for all i across all time periods being given by:

$$\frac{\partial U_i}{S_{it}}: \begin{cases} c'(S_{it}) \geq b \cdot h[A_{it} - A_{i0} - \zeta_k] \dfrac{\partial A_{it}}{\partial S_{it}} + ... + b \cdot h[A_{i,t+k-1} - A_{i,t-1} - \zeta_k] \dfrac{\partial A_{i,t+k-1}}{\partial S_{it}} \\ + b \cdot h[A_{i,t+k} - A_{it} - \zeta_k] \left(\dfrac{\partial A_{i,t+k}}{\partial S_{it}} - \dfrac{\partial A_{it}}{\partial S_{it}} \right) + ... + b \cdot h[A_{iT} - A_{i,T-k} - \zeta_k] \left(\dfrac{\partial A_{iT}}{\partial S_{it}} - \dfrac{\partial A_{i,T-k}}{\partial S_{it}} \right), & \text{for } t \leq T-k \\ c'(S_{it}) \geq b \cdot h[A_{it} - A_{i,t-k} - \zeta_k] \frac{\partial A_{it}}{\partial S_{it}} + ... + b \cdot h[A_{iT} - A_{i,T-k} - \zeta_k] \frac{\partial A_{iT}}{\partial S_{it}} & \text{for } t > T-k. \end{cases}$$

Before the necessary structure is imposed to estimate counterfactuals with the estimated structural parameters, the framework generates two insights into value-added schemes. I present these as theoretical propositions:

Proposition 4 *If ability is additively separable, then school inputs are independent of student ability.*

Proof: See Appendix A. ■

The intuition behind Proposition 4 is straightforward: since student achievement in every year is a function of ability, making achievement targets based on prior achievement eliminates student ability from the school problem. If ability is not additively separable, then school inputs are a function of student ability, but to a lesser extent than under the fixed-target scheme. In this non-additive case, a positive interaction between ability and school inputs causes schools to invest more in high-ability relative to low-ability students, exacerbating test score gaps.

Proposition 5 *If the test score improvement target $\zeta_k = \zeta \ \forall k$, student achievement is (weakly) increasing in the number of years (k) that the test score used to determine the student-specific targets is lagged.*

Proof: See Appendix A. ■

I highlight the intuition behind Proposition 5 by comparing a traditional value-added scheme ($k = 1$) to a multiperiod value-added scheme ($k = 2$). For exposition, assume that there are no dynamic complementarities: under the traditional value-added scheme, an increase in period one inputs, S_{i1}, raises the likelihood that a student exceeds the threshold

in period one. The marginal benefit of these school inputs is zero in period two, however. This is because any increase in period two achievement caused by the higher level of period one inputs is subsumed into the new period two achievement target. Multiperiod value-added schemes, in contrast, reduce this dynamic disincentive to invest. Period one inputs in this counterfactual scheme increase the likelihood that a student exceeds the threshold for periods one and two, as the target takes two periods to adjust to the higher period-one achievement level.[54] Dynamic complementarities in the production function further magnify the performance of value-added schemes with higher values of k. While the fundamental intuition for this dynamic disincentive is well-known (see Macartney (2012) for compelling evidence), the counterfactual framework provides a practical policy solution that eliminates the disincentive, using a multiperiod value-added scheme that sets student-specific targets based on *baseline* test scores.[55]

To employ the estimated structural parameters from Section 6, I introduce the two-period structure with a production technology given by Equation 6.1. The school's problem under a value-added scheme with parameter k becomes:

$$\max_{S_{i1},S_{i2}} bH(A_{i1} - A_{i0} - \zeta_k) + bH(A_{i2} - A_{i,2-k} - \zeta_k) - c(S_{i1}) - c(S_{i2})$$
$$\textit{subject to:} \qquad S_{it} \geq 0 \qquad \forall i, t\,. \tag{7.3}$$

The first-order conditions and the parametrization of the cost function yield a system of two equations, which depend on the value of k. For $k = 1$, we have:

$$\psi S_{i1}^* = bh(S_{i1}^* - \zeta_1) + bh(S_{i2}^* + \beta S_{i1}^* S_{i2}^* - \zeta_1)\,[\beta S_{i2}^*] \tag{7.4}$$

$$\psi S_{i2}^* = bh(S_{i2}^* + \beta S_{i1}^* S_{i2}^* - \zeta_1)\,[1 + \beta S_{i1}^*]\,, \tag{7.5}$$

[54] More generally, a value-added scheme with a parameter of k increases a school's payoff for period t inputs for $t - k - 1$ years.

[55] Macartney (2012) notes that this dynamic disincentive is eliminated when achievement target growth equals the growth rate of test scores. A policy that sets growth targets equal to test score growth is hard to implement, however, as the policymaker needs to know student-specific test score growth rates.

while for $k = 2$, the equations are:

$$\psi S_{i1}^* = bh(S_{i1}^* - \zeta_2) + bh(S_{i1}^* + S_{i2}^* + \beta S_{i1}^* S_{i2}^* - \zeta_2)\left[1 + \beta S_{i2}^*\right] \tag{7.6}$$

$$\psi S_{i2}^* = bh(S_{i1}^* + S_{i2}^* + \beta S_{i1}^* S_{i2}^* - \zeta_2)\left[1 + \beta S_{i1}^*\right] . \tag{7.7}$$

From Proposition 4, we know that school inputs are independent of student ability under this production technology. Given a value-added target, ζ_k, and our parameter estimates, $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$, Equations 7.4 and 7.5 (Equations 7.6 and 7.7) allow S_{i1}^* and S_{i2}^* $\forall i$ to be solved for under the value-added scheme where $k = 1$ ($k = 2$).

Under a value-added scheme, the test score improvement target, ζ_k, is set by the policymaker. A very high or very low test score improvement target reduces incentives for schools to expend resources, since the target is either too easy or too hard to attain. For the counterfactual analysis, the target, ζ_k, is assumed to be set optimally by the policymaker to maximize student achievement. In this case, setting $\zeta_1 = 0.100$ and $\zeta_2 = 0.265$ maximizes student achievement for the value-added scheme of type $k \in \{1, 2\}$, respectively.[56]

7.2 Results

The NCLB achievement level threshold in North Carolina is set at about the sixteenth percentile of the achievement distribution. Figure 8 shows the effect of moving the achievement threshold in terms of both average student achievement and the black-white test score gap. Average achievement increases as the fixed-target threshold is raised, reaching a maximum near the median of the achievement distribution (where $\delta = 0.9$). Beyond that, increases in the achievement threshold cause a decline in average achievement.

Looking at a common measure of inequality, the black-white test score gap, yields a very different picture. Initially, the achievement target is very low and so schools do not

[56]Since a unique optimal ζ_k exists for each set of structural parameters, I numerically solve for the optimal ζ_k. Intuitively, this parameter roughly maximizes the marginal benefit of investment by equating total school inputs over time period k to $\frac{\zeta_k}{k}$, making the average probability that each student exceeds the achievement target over the k periods approximately $h(0)$, maximizing the pdf $h(\cdot)$.

respond to it, making the black-white test gap identical to the 'no accountability' case. As the achievement target rises, the test gap falls to a minimum near NCLB's achievement threshold (minimized at $\delta = -0.1$). Then, as the achievement threshold continues to increase, schools shift their targeting to higher ability students, who are more likely to be white, increasing the test score gap. Finally, once the achievement target becomes too high, the policy becomes ineffectual and the test score gap returns back to the gap without accountability.[57] Therefore, a policymaker who cares about both average achievement and inequality should set the achievement threshold of a fixed target scheme within the region defined by $\delta \in [-0.1, 0.9]$.

The horizontal lines in Subfigures 8(a) and 8(b) represent average achievement and the black-white test score gap under traditional and multiperiod value-added schemes. For average achievement, we see that a multiperiod value-added scheme vastly outperforms the traditional value-added scheme by about 0.18σ, highlighting the dynamic disincentives inherent in traditional value-added schemes. Dynamic disincentives cause certain fixed-target schemes to outperform a traditional value-added scheme. Since both types of value-added scheme (represented by the horizontal line in Figure 8(b)) fare worse than the fixed-target scheme in terms of student inequality, this suggests that traditional value-added schemes can lead to lower student achievement *and* higher inequality relative to fixed-target schemes. Comparing fixed-target schemes to multiperiod value-added schemes, however, highlights an efficiency-equity trade-off: for example, replacing NCLB with a multiperiod VA scheme would increase average test scores by 0.25σ, but would also lead to a twenty percent increase (or 0.13σ) in the black-white test score gap.

Dynamic complementarities generate much of the improvement in these accountability schemes. Suppose that there is no dynamic complementarity in skill accumulation: overall inputs decline due to a decrease in the marginal benefit of investment and inputs become more concentrated in period one, since educators stop equating inputs over time in order to

[57] The figures (not shown) for the variance of the achievement distribution and the counterfactual SES test score gap (defined by the achievement difference between SES and non-SES students) are near-identical to that of the black-white test gap figure.

leverage dynamic complementarities. Without dynamic complementarities, average achievement is 0.782σ, 0.790σ and 0.830σ under NCLB, traditional value-added and multiperiod value-added schemes, respectively. If dynamic complementarities exist, but educators do not account for them in their input decisions, then average achievement is 0.803σ, 0.812σ and 0.870σ for NCLB, traditional value-added and multiperiod value-added schemes, respectively. Therefore, the key gain that dynamic complementarities provide under an incentive scheme consists in the fact that they raise the marginal benefit of investment, sharpening the incentives. Relative to a 'no accountability' world, dynamic complementarities account for about 54, 64, and 70 percent of the improvement in average achievement under NCLB, traditional value-added and multiperiod value-added schemes, respectively.

Value-added schemes that close achievement gaps: Finally, given the much higher average performance of multiperiod value-added schemes, I consider a variant of the multiperiod value-added scheme that makes the benefits of having a student exceed the threshold subgroup-specific. I focus on the black-white test gap in this example, though other test gaps can be considered: in general, this is equivalent to putting more weight on the performance of students from certain subgroups.[58]

I propose a scheme that sets benefits for black students and for all other students at 1.075 and 0.967 times NCLB's per student benefit, respectively. The test score improvement targets that maximize student performance under these benefits are 0.389 for black students and 0.312 for all other students. Costs under this multiperiod value-added scheme are the same as under the previous multiperiod value-added scheme.[59] Considering this multiperiod value-added scheme, the black-white test score gap is identical to the gap under NCLB, while average student achievement increases by 0.24σ. The reason that multiperiod value-added schemes can reduce test score gaps while maintaining a similar level of average student achievement is that while costs are convex, the benefits of investment under dynamic

[58] Alternatively, a policymaker can weight low-ability students more than high-ability students to achieve a similar effect.

[59] Costs are determined by the number of students that exceed the threshold under the accountability scheme.

complementarities are also convex.[60]

Table 4 summarizes these key statistics from the counterfactual distributions of achievement under various fixed-target and value-added schemes, as well as from a 'no accountability' world. Column (1) shows the average achievement of students, while Columns (2)-(4) show various measures of inequality, including test score gaps. While any accountability scheme increases student achievement, multiperiod value-added schemes deliver the highest level of student achievement, and can be carefully constructed to reduce inequality.

8 Discussion

Subsection 8.1 checks the validity of the structural model assumptions in Section 6, while Subsection 8.2 provides evidence that teacher effort is generating the improvement in student achievement under accountability.

8.1 Validity of Structural Assumptions

School knowledge of the technology: A key assumption in the structural estimation is that schools know the underlying production technology. While this assumption is not directly testable, I provide evidence that schools respond to the dynamic nature of the learning technology. To do so, I show that schools invest more inputs in students that are expected to be held accountable in future periods, revealing that schools internalize the cumulative nature of the learning technology – whereby investments today produce future benefits. This is a direct test of Proposition 1's corollary in Section 3.

I test for this behavior by checking if reduced-form estimates depend on grade-specific subgroup sizes. The intuition underlying this test is that schools should form expectations based on subgroup cohort size: a large period one cohort in a given subgroup increases the

[60]The global convexity arises from the multiplicative form that the dynamic complementarities take in the production technology. In practice, the convexity in the production technology is a local convexity as there is a biological upper limit to knowledge.

likelihood that the subgroup will be held accountable for the next two time periods. I therefore run grade-specific regressions that vary non-parametrically with the size of the grade three cohort, relative to the grade four and five cohorts.[61] The grade-specific regressions identify grade-specific school responses to changes in their expectations: schools should only increase inputs among students that will remain in the school (i.e. grades 3 and 4).

Appendix Figure A5 shows the relationship between grade-specific RD estimates and the relative size of a school's grade 3 cohort. Subfigures A5(a) and A5(b) show that, as the grade 3 subgroup cohort gets larger, the reduced-form estimates for periods one and two (i.e. grades 3 and 4) increase. As expected, there is no such relationship in period three (represented by Subfigure A5(c)), since these students will leave the school next year. Overall, these subfigures demonstrate that schools know and respond to the cumulative learning technology.

Expectations: In Section 6, parameters are identified structurally under the assumption that expectations are formed based on the empirical probability of treatment in the next period, given their enrollment. If schools have information unknown to the econometrician that affects the likelihood of future treatment, then the school will form a different expectation than what is assumed in the structural model. I therefore show robustness to different school expectations.

Since the structural parameters are identified off treated schools in period one, I focus on these schools. Under rational expectations, $\mathbb{E}[\Gamma_{it} = 1] = 0.60$ for these schools.[62] Appendix Figure A6 reports the estimated structural parameters and their associated confidence intervals for $\mathbb{E}[\Gamma_{it} = 1] \in [0.5, 1]$. Overall, the figures indicate that different treatment expectations affect the structural parameters in Section 6, though not substantially: if schools predict that

[61] Specifically, I run a variant of Equation 5.1 where ability is replaced with a grade 3 cohort size measure. For this measure, I use the total number of grade 3 students who belong to a subgroup divided by the total number of students belonging to that subgroup in grades 3, 4, and 5.

[62] Expectations are above fifty percent since schools with slightly more than forty students are more likely than not to receive treatment in the following year. A school's ability to foresee its period two accountability status depends upon the propensity for students to switch schools and year-to-year variation in subgroup size of the entering cohort.

they will always be treated in the next period, for instance, then the structural parameters are about forty percent higher.[63] Further, this overestimation does not substantially alter the counterfactual results in Section 7: applying structural estimates from the the case where $\mathbb{E}[\Gamma_{it}=1]=1$ yields an average achievement of 0.83σ, 0.87σ, and 1.00σ under NCLB, value-added and multiperiod value-added schemes, respectively. The respective black-white test score gaps are 0.61σ, 0.70σ and 0.70σ.

8.2 Identifying Mechanisms

While educators have many tools at their disposal for improving student achievement, this subsection provides evidence that teachers increase their effort in response to accountability pressure. To do so, I show that teacher effectiveness is related to the proportion of students in the accountable subgroup within the teacher's class, independently of teacher effectiveness in prior years. I also find little evidence of within-class spillovers across students, supporting the modelling assumption that inputs are student-specific. Finally, I rule out several alternative methods that schools may use to increase student achievement.

The methodology – which I develop formally in Appendix Section B – builds upon Macartney et al. (2016) and relates teacher effectiveness (measured by value-added) to the proportion of accountable students in her classroom. The idea is that if teacher effort drives the subgroup accountability effect, then teacher effort should increase with the number of accountable students in the class. Keeping with the spirit of the RD design, I only investigate teachers in schools with around forty students in the accountable subgroup. To account for any systematic sorting of accountable students to teachers (which I explicitly test for below), teacher ability and baseline effort levels (measured by teacher value-added in other years) are controlled for implicitly.

Appendix Figure A7 shows the relationship between teacher effectiveness and the proportion of students in the relevant subgroup for schools that lie just to the left of the discontinuity

[63] When $\mathbb{E}[\Gamma_{it}=1]=1$, the estimated $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$ are 14.52 (s.e. 3.50) and 14.51 (s.e. 2.18), respectively.

(not facing subgroup-specific accountability) and those just to the right of the discontinuity (facing subgroup-specific accountability). As expected, there is no relationship to the left, but a strong and statistically significant relationship to the right of the discontinuity. These results indicate that, even once ability is controlled for, teachers who have more students from the accountable subgroup in their class are more effective. Furthermore, this relationship only exists for teachers in school-years where these students are held accountable, indicating that teachers increase effort as more accountable students enter their class.[64]

Next, to investigate whether teacher effort is student- or class-specific, I search for within-class spillovers across subgroups. These types of spillover would indicate that teachers cannot specifically target their increased effort towards accountable students,[65] violating the modelling assumption that school inputs are student-specific. To explore such spillovers, I investigate the relationship between the proportion of accountable students in a class and the achievement of non-accountable students. If these non-accountable students receive a test score boost when the proportion of accountable students (and teacher effort) rises, then a classroom-level model of effort is more appropriate. Appendix Table A6 reports the relationship between the number of accountable students in a class and test scores for students belonging versus not belonging to the accountable subgroup.[66] Column (1) shows that there is no relationship between the proportion of accountable students in a class and the achievement of non-accountable students for schools with either more or less than forty students. This lends support to a student-specific model of teacher effort.

Alternative Mechanisms: Two alternative mechanisms that could lead to subgroup-specific improvements in test scores that are unrelated to teacher effort involve student sorting and resource targeting, respectively. I test for student sorting – where accountable students are assigned to classrooms with higher-achieving peers or better teachers – by using

[64]While this method rules out sorting as a possible mechanism, there remain some plausible alternative mechanisms that may generate this relationship. For example, such a relationship could appear if principals hire tutors for each accountable student.

[65]Alternatively, it could imply large within-class peer effects.

[66]A strong positive (negative) relationship for students belonging to the subgroup is indicative of either increasing (decreasing) returns to scale in student-specific effort or within subgroup peer effects.

the regression discontinuity design from Section 4 and replacing test scores with measures of teacher and peer quality. I measure peer quality using classmate achievement levels from the prior year to avoid the reflection problem (Manski, 1993). For teacher quality, leave-out teacher value-added estimates are computed using the methodology described in Chetty et al. (2014). I search for evidence of resource targeting by checking for class size changes around the discontinuity.

Appendix Figure A8 shows a visual representation of the reduced-form results. It is clear that there are no jumps in peer quality, teacher quality, or class size for students belonging or not belonging to the accountable subgroup once the forty student threshold is crossed. Therefore, there is no evidence of student sorting or resource targeting in response to subgroup-specific accountability.

9 Conclusion

In this paper, I have developed a new methodology for cleanly identifying dynamic interactions among school inputs for the first time. The methodology adapts a regression discontinuity design to incorporate year-to-year treatment variation, providing the effective period-by-period randomization necessary to identify these dynamic complementarities. I apply the methodology to an education context, where a dynamic perspective is natural given the cumulative nature of the underlying education process. Using a rule that certain students are held accountable only if their subgroup has forty students or more at the school level, I find that dynamic complementarities are important and that schools take the underlying skill accumulation technology into account when making their period-specific input choices.

The results shed new light on incentive design in education by emphasizing the role of dynamic complementarities. Following "A Nation at Risk" (United States National Commission on Excellence in Education, 1983), a damning report that brought to light the parlous

state of public education in the US, policymakers have made extensive use of incentive schemes. Without an adequate understanding of the production function for educational achievement, however, policymakers cannot predict the full dynamic impact of alternative accountability schemes in a reliable way. By identifying the underlying dynamic technology, I am able to build a framework that provides policymakers with the full distribution of student achievement under alternative schemes that have yet to be implemented. In particular, I propose a feasible incentive scheme that both raises student achievement and reduces inequality.

In other education settings, the ability to identify dynamic complementarities opens up a host of policy design and evaluation issues that should lead to more effective interventions. Dynamic complementarities imply that policymakers should implement interventions that affect children over multiple periods, rather than focusing on a narrow developmental stage. Similarly, education reforms should, ideally, be evaluated over long time horizons (rather than a simple before and after comparison) to allow these dynamic complementarities to be leveraged by educators. Furthermore, programs targeting a narrow developmental stage should be evaluated differently among those receiving relatively high versus low inputs in future periods so that policymakers can better target future interventions, as in the analysis of Head Start by Currie and Thomas (2000).

Since dynamic complementarities exist in other settings, the broader significance of the research is not limited to education. For example, in related work, I am exploring the implications of dynamic complementarities in terms of whether an incentive scheme should be implemented at the group or individual level. Under dynamic complementarities, individuals must match their inputs across time, which increases the importance of coordination across individuals, thus favouring group-level incentives that encourage coordination.

References

Agostinelli, Francesco and Matthew Wiswall (2016), "Estimating the technology of children's skill formation." Working Paper 22442, National Bureau of Economic Research, URL http://www.nber.org/papers/w22442.

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.

Aizer, Anna and Flávio Cunha (2012), "The production of human capital: Endowments, investments and fertility." Working Paper 18429, National Bureau of Economic Research, URL http://www.nber.org/papers/w18429.

Almond, Douglas and Bhashkar Mazumder (2013), "Fetal origins and parental responses." *Annual Review of Economics*, 5, 36–56.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters (2012), "Who benefits from KIPP?" *Journal of policy Analysis and Management*, 31, 837–860.

Chakrabarti, Rajashri (2014), "Incentives and responses under No Child Left Behind: Credible threats and the role of competition." *Journal of Public Economics*, 110, 124 – 146.

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

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz (2016), "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment." *American Economic Review*, 106, 855–902.

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.

Cunha, Flávio and James Heckman (2007), "The technology of skill formation." *American Economic Review*, 97, 31–47.

Cunha, Flávio and James J. Heckman (2008), "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of Human Resources*, 43, 738–782.

Cunha, Flávio, James J. Heckman, and Susanne M. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78, 883–931.

Currie, Janet and Duncan Thomas (2000), "School quality and the longer-term effects of Head Start." *Journal of Human Resources*, 35, 755–74.

Dee, Thomas S. and Brian Jacob (2011), "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management*, 30, 418–446.

Dell, Melissa (2010), "The persistent effects of Peru's mining mita." *Econometrica*, 78, 1863–1903.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks (2016), "School accountability, postsecondary attainment and earnings." *Review of Economics and Statistics*, 98, 848–862.

Ding, Weili and Steven F. Lehrer (2010), "Estimating treatment effects from contaminated multiperiod education experiments: The dynamic impacts of class size reductions." *Review of Economics and Statistics*, 92, 31–42.

Farber, Matthew S. (2016), *Essays on the Economics of Education of Underserved Populations*. Ph.D. thesis, University of Texas at Austin.

Figlio, David and Susanna Loeb (2011), "School accountability." In *Handbook of the Economics of Education* (Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds.), volume 3, 383 – 421, Elsevier.

Figlio, David N. (2006), "Testing, crime and punishment." *Journal of Public Economics*, 90, 837–851.

Figlio, David N. and Joshua Winicki (2005), "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics*, 89, 381–394.

Gaddis, S. Michael and Douglas Lee Lauen (2014), "School accountability and the black–white test score gap." *Social Science Research*, 44, 15–31.

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano (2016), "Do fiscal rules matter?" *American Economic Journal: Applied Economics*, 8, 1–30.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz (2010), "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program." *Quantitative Economics*, 1, 1–46.

Jacob, Brian A. (2005), "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics*, 89, 761–796.

Johnson, Rucker C. and C. Kirabo Jackson (2017), "Reducing inequality through dynamic complementarity: Evidence from head start and public school spending." Working Paper 23489, National Bureau of Economic Research, URL `http://www.nber.org/papers/w23489`.

Koretz, Daniel M. (2008), *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press, Cambridge, MA.

Krieg, John M. (2008), "Are students left behind? The distributional effects of the No Child Left Behind Act." *Education Finance and Policy*, 3, 250–281.

Krieg, John M. (2011), "Which students are left behind? The racial impacts of the No Child Left Behind Act." *Economics of Education Review*, 30, 654–664.

Krueger, Alan B. (1998), "Reassessing the view that American schools are broken." *Economic Policy Review*, 4, 29–43.

Krueger, Alan B. (2003), "Economic considerations and class size." *Economic Journal*, 113, F34–F63.

Lauen, Douglas Lee and S. Michael Gaddis (2012), "Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability on student achievement." *Educational Evaluation and Policy Analysis*, 34, 185–208.

Lee, David S. and David Card (2008), "Regression discontinuity inference with specification error." *Journal of Econometrics*, 142, 655–674.

Lee, David S. and Thomas Lemieux (2010), "Regression discontinuity designs in economics." *Journal of Economic Literature*, 48, 281–355.

Lubotsky, Darren and Robert Kaestner (2016), "Do 'skills beget skills'? Evidence on the effect of kindergarten entrance age on the evolution of cognitive and non-cognitive skill gaps in childhood." *Economics of Education Review*, 53, 194 – 206.

Macartney, Hugh (2012), *The Dynamic Effects of Educational Accountability*. Ph.D. thesis, University of Toronto.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2015), "Incentive design in education: An empirical analysis." Working Paper 21835, National Bureau of Economic Research, URL `http://www.nber.org/papers/w21835`.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2016), "A unifying framework for education policy analysis.", URL `http://homes.chass.utoronto.ca/~mcmillan/MMP3_6nov16a.pdf`.

Malamud, Ofer, Cristian Pop-Eleches, and Miguel Urquiola (2016), "Interactions between family and school environments: Evidence on dynamic complementarities?" Working Paper 22112, National Bureau of Economic Research, URL http://www.nber.org/papers/w22112.

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.

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.

Papay, John P., Richard J. Murnane, and John B. Willett (2014), "High-school exit examinations and the schooling decisions of teenagers: Evidence from regression-discontinuity approaches." *Journal of Research on Educational Effectiveness*, 7, 1–27.

Porter, Kristin E., Sean F. Reardon, Fatih Unlu, Howard S. Bloom, and Joseph R. Cimpian (2017), "Estimating causal effects of education interventions using a two-rating regression discontinuity design: Lessons from a simulation study and an application." *Journal of Research on Educational Effectiveness*, 10, 138–167.

Reback, Randall (2008), "Teaching to the rating: School accountability and the distribution of student achievement." *Journal of Public Economics*, 92, 1394–1415.

Sims, David P. (2013), "Can failure succeed? Using racial subgroup rules to analyze the effect of school accountability failure on student performance." *Economics of Education Review*, 32, 262–274.

Spellings, Margaret (2005), "No Child Left Behind: A road map to state implementation." Technical report, US Department of Education, URL `http://www.ed.gov/admins/lead/account/roadmap/roadmap.pdf`.

Springer, Matthew G. (2008), "The influence of an NCLB accountability plan on the distribution of student test score gains." *Economics of Education Review*, 27, 556–563.

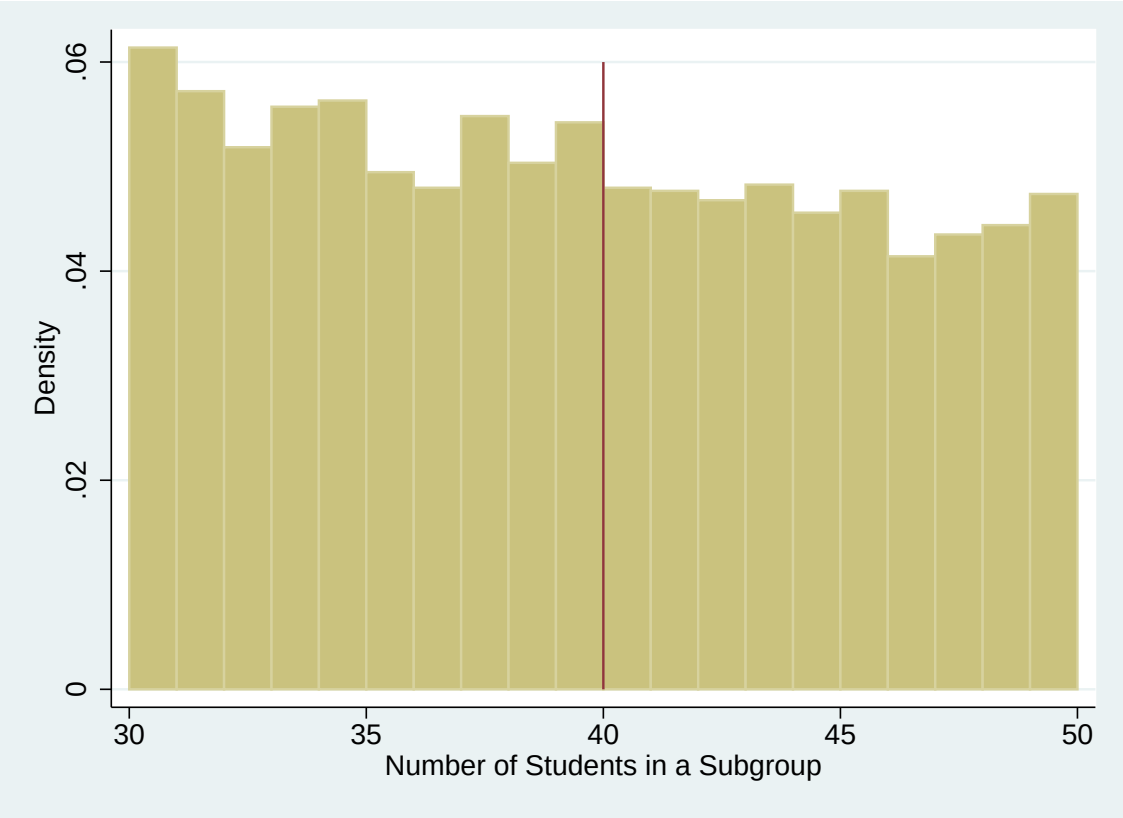
Stullich, Stephanie, Elizabeth Eisner, Joseph McCrary, and Collette Roney (2006), "National assessment of Title I: Interim report, Vol. I: Implementation of Title I." Technical report, Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Todd, Petra E. and Kenneth I. Wolpin (2003), "On the specification and estimation of the production function for cognitive achievement." *Economic Journal*, 113, F3–F33.

Todd, Petra E. and Kenneth I. Wolpin (2007), "The production of cognitive achievement in children: Home, school, and racial test score gaps." *Journal of Human Capital*, 1, 91–136.

United States National Commission on Excellence in Education (1983), *A nation at risk: The Imperative for Educational Reform: A Report to the Nation and the Secretary of Education, United States Department of Education*. Washington, D.C.: The Commission: [Supt. of Docs., U.S. G.P.O. distributor].

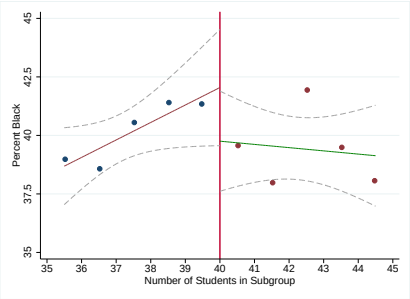
Figure 2: Density of the Running Variable



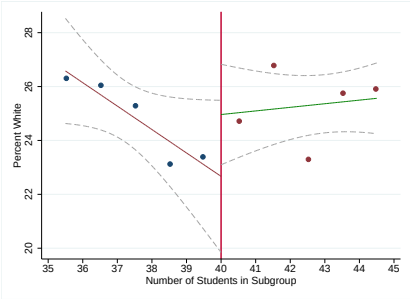
Notes: Figure 2 is based on 3,356 observations. The vertical line indicates the forty-student threshold. The p-value from the McCrary (2008) test is computed using a bandwidth of 5 and a bin width of 1.

Figure 3: Covariates (Students belong to Subgroup)

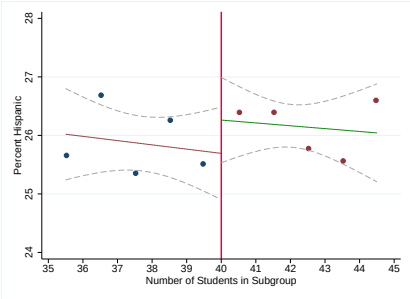
(a) Student is Black
RD Estimate: -2.30 (1.61)



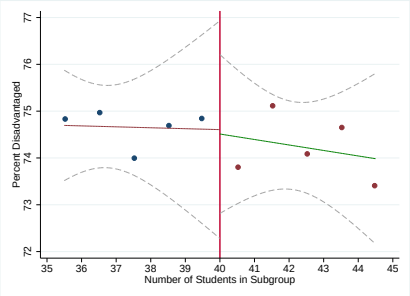
(b) Student is White
RD Estimate: 2.30 (1.67)



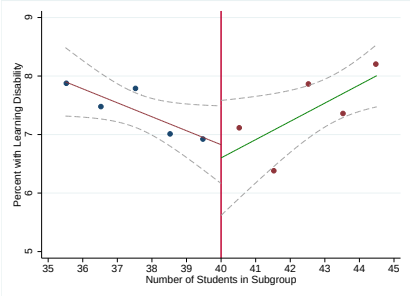
(c) Student is Hispanic
RD Estimate: 0.57 (0.52)



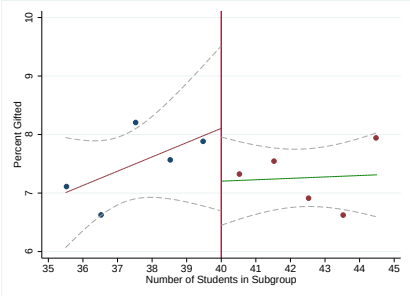
(d) Student is Disadvantaged
RD Estimate: -0.09 (1.42)



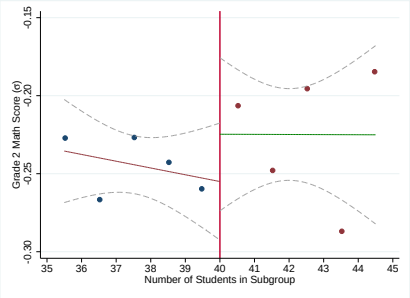
(e) Student has a Learning Disability
RD Estimate: -0.23 (0.57)



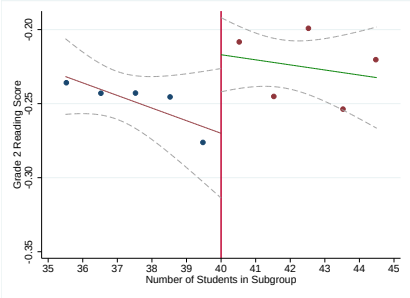
(f) Student is Gifted
RD Estimate: -0.91 (0.79)



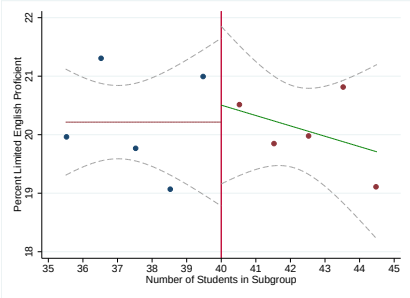
(g) Grade 2 Math Score
RD Estimate: 0.030 (0.30)



(h) Grade 2 Reading Score
RD Estimate: 0.048* (0.025)



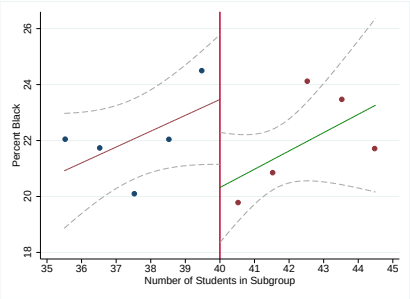
(i) Student is an English Learner
RD Estimate: 0.29 (0.98)



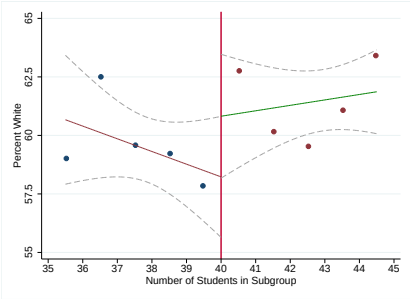
Notes: All figures are based on 1,655 observations. Figures and RD estimates control for subgroup fixed-effects. Each RD estimate comes from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. RD point estimates correspond to those in Panel A of Appendix Table A1. Dashed lines represent 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 4: Covariates (Students do not belong to Subgroup)

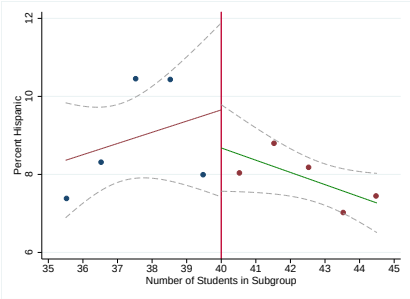
(a) Student is Black
RD Estimate: –2.17* (1.11)



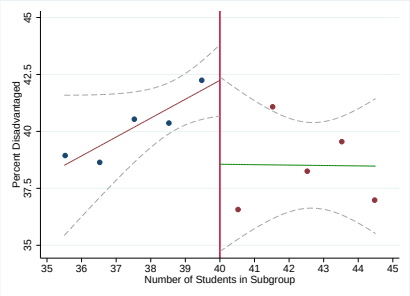
(b) Student is White
RD Estimate: 2.14 (1.52)



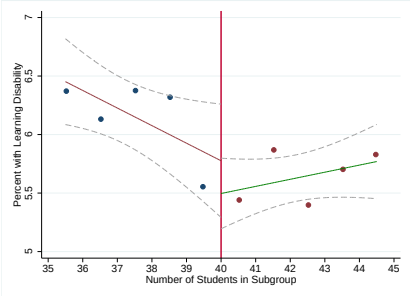
(c) Student is Hispanic
RD Estimate: -0.72 (0.86)



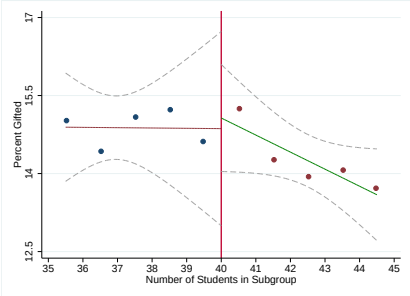
(d) Student is Disadvantaged
RD Estimate: -2.99* (1.59)



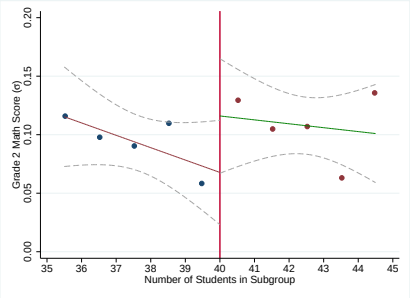
(e) Student has a Learning Disability
RD Estimate: -0.28 (0.28)



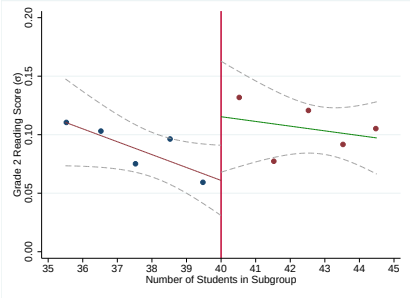
(f) Student is Gifted
RD Estimate: 0.21 (1.03)



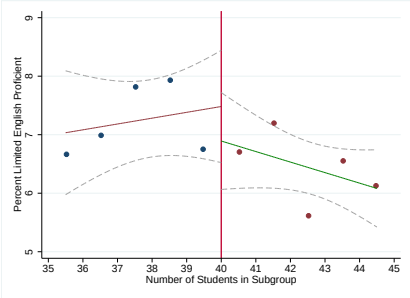
(g) Grade 2 Math Score
RD Estimate: 0.048 (0.032)



(h) Grade 2 Reading Score
RD Estimate: 0.047* (0.027)

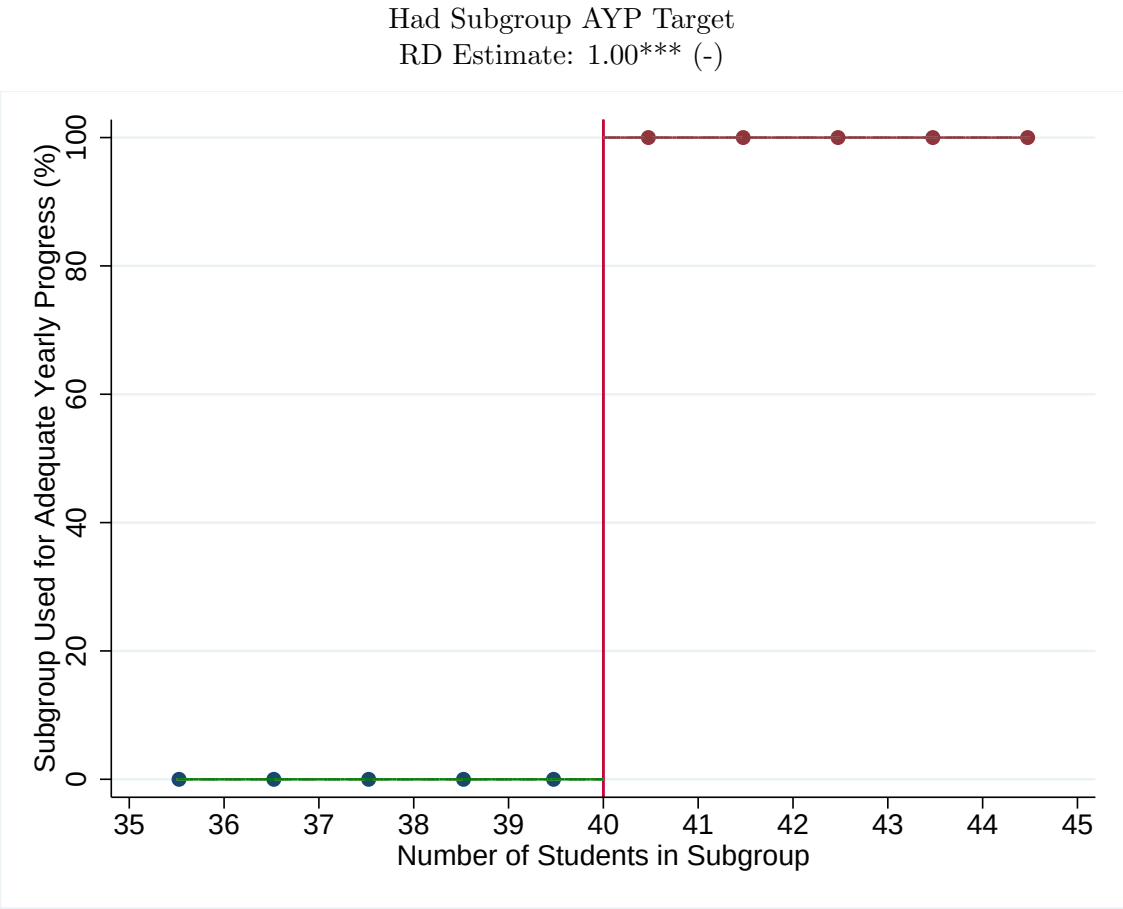


(i) Student is an English Learner
RD Estimate: -0.59 (0.62)



Notes: All figures are based on 1,655 observations. Figures and RD estimates control for subgroup fixed-effects. Each RD estimate comes from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. RD point estimates correspond to those in Panel B of Appendix Table A1. Dashed lines represent 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 5: First Stage (Faced Subgroup Accountability Pressure)

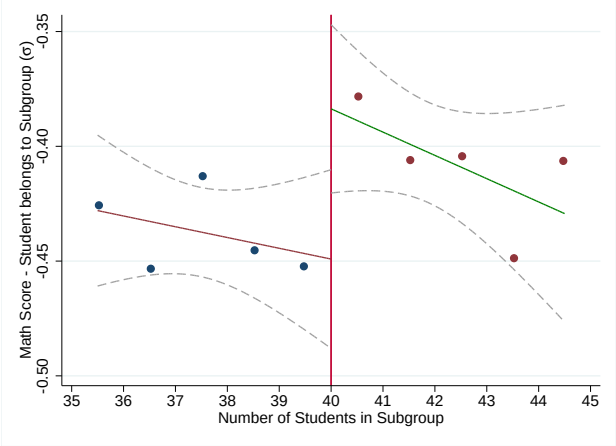


Notes: Figure 5 is based on 1,655 observations. The x-axis represents the number of students taking the end-of-grade test and who are held accountable under NCLB (see Section 2 for details). The RD estimate has no standard errors as the line is perfectly fitted. Significance levels: *** 1 percent.

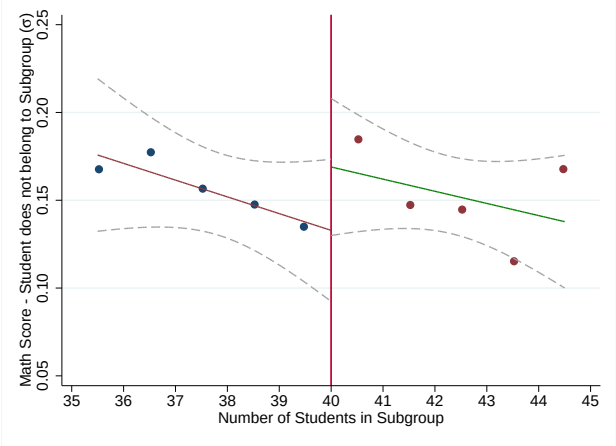
Figure 6: Reduced-Form Relationship between Test Scores and Number of Students in a Subgroup

Math Scores

(a) Students belong to Subgroup
RD Estimate: 0.066** (0.026)

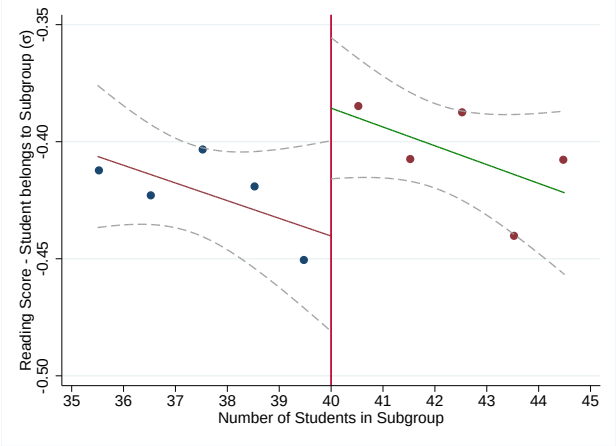


(b) Students do not belong to Subgroup
RD Estimate: 0.037 (0.027)

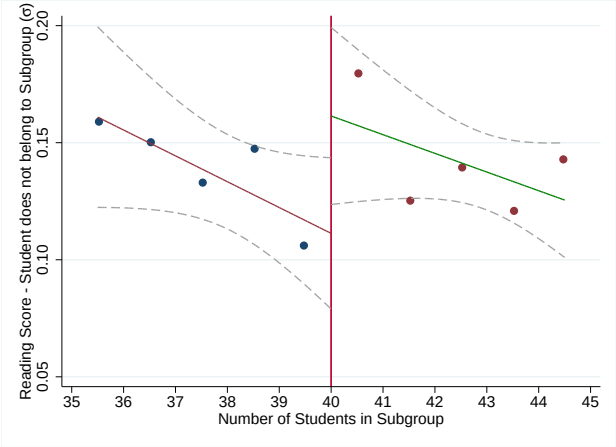


Reading Scores

(c) Students belong to Subgroup
RD Estimate: 0.055** (0.025)



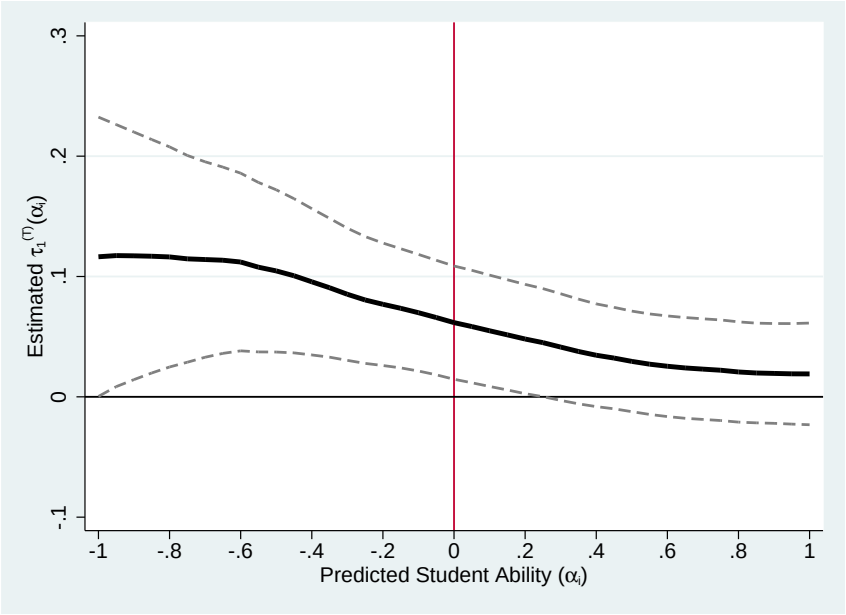
(d) Students do not belong to Subgroup
RD Estimate: 0.051** (0.024)



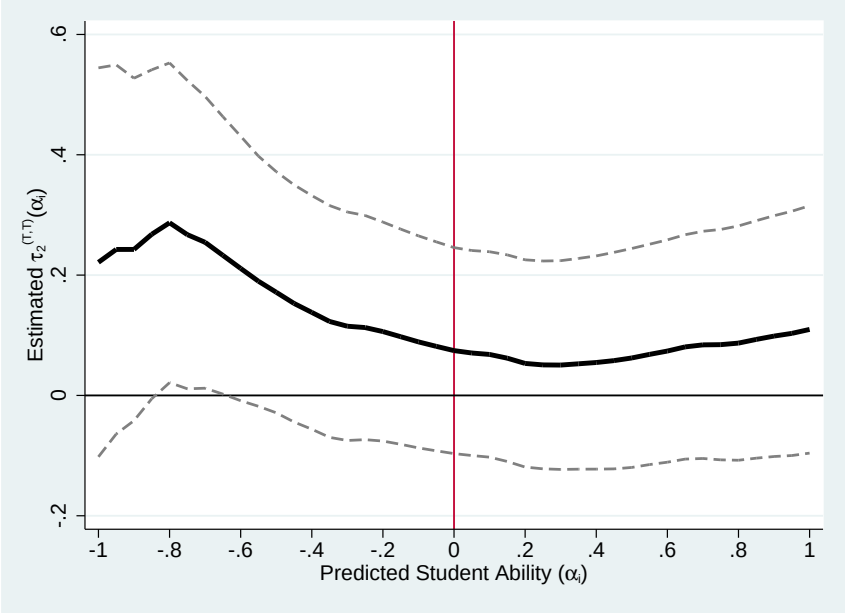
Notes: Figures are based on 1,655 observations. Figures and RD estimates control for subgroup and year fixed effects since the large differences in average test scores across subgroups (see Appendix Table A3) lead to high standard errors without their inclusion. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 7: Estimated Treatment Effects by Predicted Ability ($\hat{\alpha}_i$)

(a) Grade 3 ($\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$)



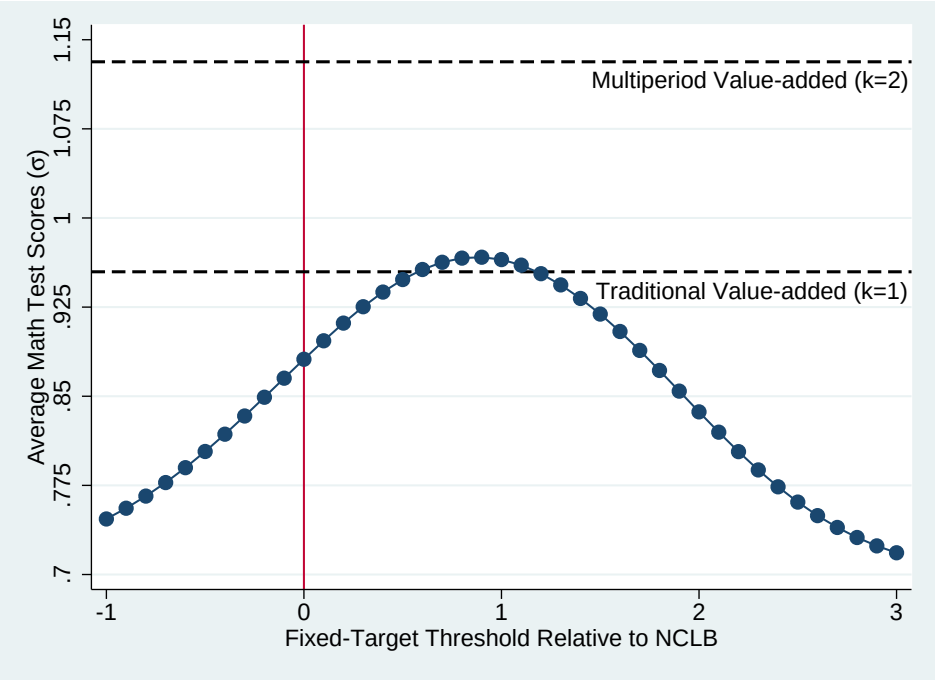
(b) Grade 4 ($\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$)



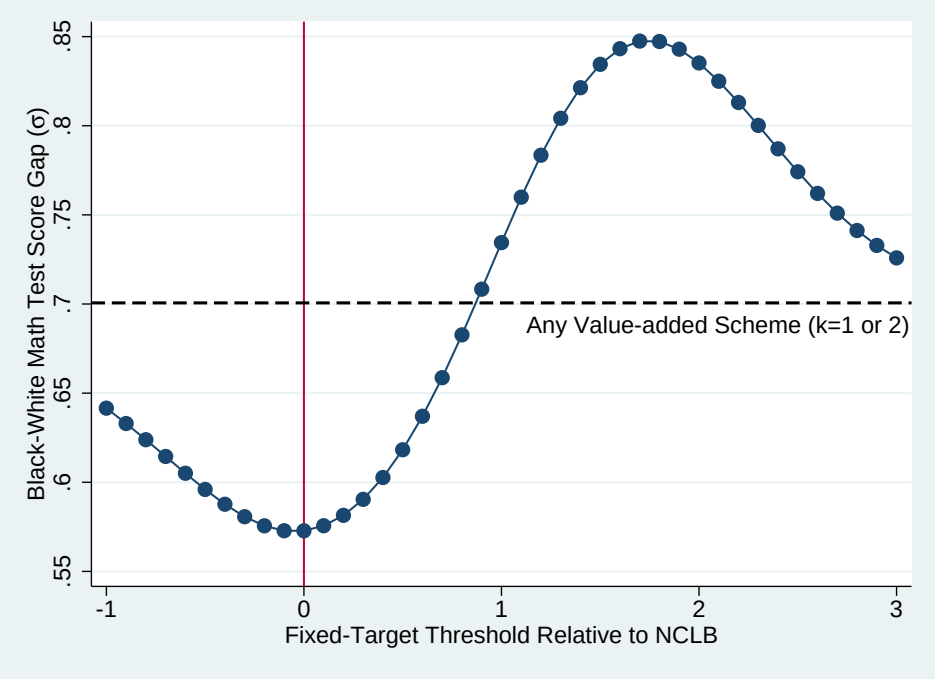
Notes: The figures plot $\hat{\tau}_1^{(T)}(\hat{\alpha}_i)$ and $\hat{\tau}_2^{(T,T)}(\hat{\alpha}_i)$ for various predicted ability levels, $\hat{\alpha}_i$. These estimates are used in the structural estimation in Section 6. Student ability is predicted using grade 2 test scores and a full set of demographic characteristics include gender, ethnicity, gifted status, English learner status, and free lunch status. The vertical line denotes estimates for students with predicted ability at the achievement threshold ($\mathcal{A}^*$). Dashed lines represent 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008).

Figure 8: Counterfactuals: Moving the Achievement Threshold

(a) Average Achievement



(b) Black-White Test Score Gap



Notes: Figure 8(a) plots the average achievement level and Figure 8(b), the black-white test score gap for various achievement thresholds. The horizontal lines represent the average achievement level or black-white test score gap under alternative value-added schemes for reference. The vertical line represents the current achievement threshold under NCLB. Table 4 displays the average achievement and various inequality measures for fixed-target schemes with the NCLB achievement target ($\delta = 0$) and the achievement target that maximizes average achievement ($\delta = 0.9$).

Table 1: Summary Statistics

	Full Sample[1] (1)	RD Sample (+/– 5)[2] (2)	RD Sample: Students in Subgroup[3] (3)	RD Sample: Students not in Subgroup[3] (4)
Mean (S.D.) of Student Characteristics				
Math Score (σ)	-0.036	-0.025	-0.188	0.137
(Student-level s.d.)	(0.492)	(0.453)	(0.392)	(0.452)
Reading Score (σ)	-0.054	-0.053	-0.222	0.117
(Student-level s.d.)	(0.467)	(0.428)	(0.362)	(0.421)
% White	37.1	42.7	25.1	60.4
	(39.5)	(40.6)	(39.6)	(33.3)
% Black	25.7	31.0	39.9	22.1
	(34.1)	(39.3)	(46.0)	(28.6)
% Hispanic	14.2	17.1	25.9	8.4
	(28.0)	(32.3)	(41.9)	(13.3)
% Asian	7.4	3.7	4.8	2.7
	(23.4)	(14.8)	(20.4)	(4.4)
% Free or Reduced	54.7	56.9	74.4	39.4
Price Lunch	(33.3)	(32.7)	(24.4)	(30.5)
% Labeled	11.2	11.1	7.4	14.7
as Gifted	(13.8)	(12.0)	(10.2)	(12.5)
% of Students	6.0	6.6	7.4	5.9
with Disability	(7.9)	(5.1)	(6.0)	(3.8)
% English	11.7	13.4	20.1	6.8
Learners	(21.4)	(23.9)	(30.9)	(9.8)
Observations	84,151	3,310	1,655	1,655
(school-subgroup-belong-year)[4]				

[1] The full sample consists of all schools whose highest grade is either 5 or 6 and who have at least twenty-five percent of students in the accountable subgroup take the test (less than one percent of sample).
[2] The RD sample is restricted to schools with a subgroup that is +/- 5 students away from the forty student threshold.
[3] 'Students in subgroup' are students who belong to the subgroup that is near the forty student threshold, while 'students not in subgroup' are students in the same school but who do belong to the subgroup that is near the forty student threshold.
[4] For each subgroup-school-year combination, there are two separate observations: students who belong and students who do not belong to the subgroup that is being analyzed.

Table 2: Regression-Discontinuity Estimates of Subgroup-Specific Accountability on Student Achievement

	Math Scores (σ)		*Reading Scores (σ)*	
	No Covariates	Covariates	No Covariates	Covariates
	(1)	(2)	(3)	(4)
Panel A. Students in Subgroup				
$\tau_{i \in g}$	0.066**	0.053**	0.055**	0.028*
	(0.026)	(0.021)	(0.025)	(0.017)
Panel B. Students Not in Subgroup				
$\tau_{i \notin g}$	0.037	0.008	0.051**	0.013
	(0.027)	(0.018)	(0.024)	(0.009)
Panel C. Difference-in-Discontinuity				
τ_{diff}	0.030	0.046***	0.004	0.015
	(0.029)	(0.016)	(0.030)	(0.017)
Observations	3,310	3,292	3,310	3,292

Notes: Number of observations are given for Panel C: Panel A and B have half that number of observations each. Grade 3 results are missing for the 2005-06 school year due to missing data on the math pre-test for that year. The bandwidth used is five. All columns include subgroup and year fixed-effects. Covariates include controls for grade 2 test scores, gender, grade, ethnicity, gifted status, English learner status, free lunch status, and grade configuration, subgroup and year fixed-effects. Standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 3: Estimates of Subgroup-Specific Accountability on Student Achievement by Grade and Prior Treatment Status

Outcome: Math Scores (σ) Prior Period Treatment Status:	Treated $\tau_t^{(T,T)}$	Untreated $\tau_t^{(U,T)}$	Difference $\tau_t^{(T,T)} - \tau_t^{(U,T)}$
Panel A. Period 1 (i.e., Grade 3)			
τ_1 ('Placebo')	0.192** (0.084)	0.164* (0.084)	0.028 (0.119)
Panel B. Period 2 (i.e., Grade 4)			
τ_2	0.320*** (0.079)	0.132* (0.075)	0.189* (0.109)
Panel C. Period 3 (i.e., Grade 5)			
τ_3	0.069 (0.108)	-0.103 (0.086)	0.172 (0.138)
Panel D. **Test for Dynamic Complementarities:**			
$\frac{1}{2}\left(\tau_2^{(T,T)} - \tau_2^{(U,T)} + \tau_3^{(T,T)} - \tau_3^{(U,T)}\right)$	0.180** (0.089)		
Covariates	Yes	Yes	Yes
Observations in Panel A	161	162	323
Observations in Panel B	211	229	440
Observations in Panel C	206	202	408

Notes: Observations are at the school-subgroup-year level and include only students who belong to the accountable subgroup. The 2002-03 school year is omitted since estimates for grades 4 and 5 must condition on the prior year treatment status. Grade 3 results are missing for the 2005-06 school year due to missing data on the math pre-test for that year. The bandwidth used is five. Covariates include controls for prior test scores, gender, ethnicity, gifted status, English learner status, free lunch status, and grade configuration, subgroup and year fixed-effects. Standard errors are clustered on student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 4: Student Achievement under Counterfactual Policies

Outcome: Math Scores (σ)				
Accountability Scheme	Average Achievement (1)	S.D. of Achievement (2)	Black-White Test Score Gap (3)	SES Test Score Gap (4)
No Child Left Behind Current Threshold ($\delta = 0$)	0.88	1.03	0.57	0.53
Counterfactuals				
No Accountability Scheme	0.70	1.10	0.70	0.64
Fixed-Target Higher Threshold ($\delta = 0.9$)	0.97	1.11	0.71	0.64
Traditional Value-Added ($k = 1$, $\zeta_1 = 0.148$)	0.95	1.10	0.70	0.64
Multiperiod Value-Added ($k = 2$, $\zeta_2 = 0.385$)	1.13	1.10	0.70	0.64
Multiperiod Value-Added, with NCLB Black-White Gap ($k = 2$, $b_b = 1.075$, $b_w = 0.967$, $\zeta_2^b = 0.389$, $\zeta_2^w = 0.312$)	1.12	1.08	0.57	0.60

Notes: All counterfactuals rely on the structural estimates from column (1) of Appendix Table A4 (see Section 7 for more detail). For column (1) only, the achievement measure is in standard deviation units that are relative to the current NCLB achievement threshold. These can be put in terms of the current test score distribution by subtracting 0.70σ. For columns (2)-(4) the inequality measures are in math score standard deviation units. The SES test score gap is the average difference in test scores between SES and non-SES students.

A Proofs

Proposition 1 *If schools invest optimally, $S_{it} \geq S_{i,t+1} \geq ... \geq S_{iT}$, $\forall i$ in expectation.*

Proof: Consider student i with the same realization of Γ_{it} and $\Gamma_{i,t+1}$ in periods t and $t+1$. The result then follows directly from the FOCs w.r.t. to S_{it} and $S_{i,t+1}$, $\forall i$:

$$\frac{\partial U_i}{S_{it}} : \Gamma_{it} \cdot b \cdot h[A_{it} - \mathcal{A}^*] \frac{\partial f_t(\cdot)}{\partial S_{it}} + p_i \cdot b \cdot h[A_{i,t+1} - \mathcal{A}^*] \frac{\partial f_{t+1}(\cdot)}{\partial S_{it}} + ... + p_i \cdot b \cdot h[A_{iT} - \mathcal{A}^*] \frac{\partial f_T(\cdot)}{\partial S_{it}} \leq c'(S_{it}) \quad \text{(A.1)}$$

$$\frac{\partial U_i}{S_{i,t+1}} : \Gamma_{i,t+1} \cdot b \cdot h[A_{i,t+1} - \mathcal{A}^*] \frac{\partial f_{t+1}(\cdot)}{\partial S_{i,t+1}} + ... + p_i \cdot b \cdot h[A_{iT} - \mathcal{A}^*] \frac{\partial f_T(\cdot)}{\partial S_{i,t+1}} \leq c'(S_{i,t+1}) \,. \quad \text{(A.2)}$$

Suppose $S_{i,t+1} > S_{it}$. Then, the school can raise S_{it} and lower $S_{i,t+1}$ (up to the point $S_{it} = S_{i,t+1}$) and gain equal or higher benefits in expectation because $\frac{\partial f_t(\cdot)}{\partial S_{it}} \geq \frac{\partial f_{t+1}(\cdot)}{\partial S_{i,t+1}}$ (by assumption) and the first term in Equation A.1 is non-negative (since $b \geq 0$, $p_i \geq 0$, $h[\cdot] \geq 0$). Since this change also reduces overall costs (since $c(\cdot)$ is convex and $S_{i,t+1} > S_{it}$), it follows that setting $S_{i,t+1} > S_{it}$ is suboptimal and therefore it must be that $S_{it} \geq S_{i,t+1}$. ■

Proposition 2 *Within a subgroup, school inputs are highest for some student with innate ability below the achievement threshold.*

Proof: Let $\delta > 0$ be an arbitrarily small constant. Consider students i and j in year t such that $p_i = p_j$, $\Gamma_{it} = \Gamma_{jt}$, and $\alpha_j = \alpha_i + \delta$ and assume that $\alpha_j \geq \mathcal{A}^*$. From Equation 3.6 for period t, we have that:

$$\begin{aligned}
&c'(S_{it}) = \Gamma_{it} \cdot b \cdot h[f_t(\alpha_i, S_{i1}, ..., S_{it}) - \mathcal{A}^*] \frac{\partial f_t(\cdot)}{\partial S_{it}} + ... + p_i \cdot b \cdot h[f_T(\alpha_i, S_{i1}, ..., S_{iT}) - \mathcal{A}^*] \frac{\partial f_T(\cdot)}{\partial S_{it}} \\
&\geq \Gamma_{jt} \cdot b \cdot h[f_t(\alpha_j, S_{j1}, ..., S_{jt}) - \mathcal{A}^*] \frac{\partial f_t(\cdot)}{\partial S_{jt}} + ... + p_j \cdot b \cdot h[f_T(\alpha_j, S_{jt}, ..., S_{jT}) - \mathcal{A}^*] \frac{\partial f_T(\cdot)}{\partial S_{jt}} = c'(S_{jt}) \\
&\therefore c'(S_{it}) \geq c'(S_{jt}) \implies S_{it} \geq S_{jt} \,,
\end{aligned}$$

where the inequality follows because (i) $\delta > 0$ so it must be that $f_t(\alpha_i, S_{i1}, ..., S_{it}) - \mathcal{A}^* \leq f_t(\alpha_i + \delta, S_{i1}, ..., S_{it}) - \mathcal{A}^*$ $\forall t$ and (ii) because the returns to school inputs are non-negative ($\frac{\partial f_t(\cdot)}{\partial S_{it}} \geq 0$ $\forall t$), decreasing in ability ($\frac{f_t(\cdot)}{\partial S_{it} \partial \alpha_i} \leq 0$) and $\alpha_i + \delta \geq \mathcal{A}^*$ (where δ is small), we have that $h(f_t(\alpha_i, S_{i1}, ..., S_{it}) - \mathcal{A}^*) \leq h(f_t(\alpha_j, S_{j1}, ..., S_{jt}) - \mathcal{A}^*)$ $\forall t$ since the error term is symmetric, unimodal and mean zero. ■

Proposition 3 $\tau_2^{(T,T)} > \tau_2^{(U,T)}$ *iff* $\beta_{12} > 0$.

Proof: Subtracting Equation 4.5 from 4.6 gives:

$$\tau_2^{(T,T)} - \tau_2^{(U,T)} = (\bar{S}_2^{(T,T)} - \bar{S}_2^{(T,U)}) - (\bar{S}_2^{(U,T)} - \bar{S}_2^{(U,U)}) + \frac{1}{N}\sum_i^N \beta_{12}\Big[S_{i1}^{(T)}(S_{i2}^{(T,T)} - S_{i2}^{(T,U)}) - S_{i1}^{(U)}(S_{i2}^{(U,T)} - S_{i2}^{(U,U)})\Big] \quad \text{(A.3)}$$

Since achievement is increasing in school inputs, we have that $S_{i2}^{(T,T)} > S_{i2}^{(T,U)}$, $S_{i2}^{(U,T)} > S_{i2}^{(U,U)}$, $S_{i2}^{(T,U)} > S_{i2}^{(U,U)}$ and $S_{i1}^{(T)} > S_{i1}^{(U)}$. Now, suppose $\beta_{12} > 0$: then, the FOCs (given by Equation 3.6) from the school problem require that $S_{i2}^{(T,T)} - S_{i2}^{(T,U)} > S_{i2}^{(U,T)} - S_{i2}^{(U,U)}$ $\forall i$ since $\frac{\partial f_2(\cdot)}{\partial S_{i2}^{(T,T)}} > \frac{\partial f_2(\cdot)}{S_{i2}^{(U,T)}}$. Since $(\bar{S}_2^{(T,T)} - \bar{S}_2^{(T,U)}) - (\bar{S}_2^{(U,T)} - \bar{S}_2^{(U,U)}) > 0$, both the first and second line of Equation A.3 are positive and so we have that $\tau_2^{(T,T)} > \tau_2^{(U,T)}$. Similarly, $\tau_2^{(T,T)} > \tau_2^{(U,T)}$ implies that $\beta_{12} > 0$, because if $\beta_{12} \leq 0$ then by the school problem $S_{i2}^{(T,T)} - S_{i2}^{(T,U)} \leq S_{i2}^{(U,T)} - S_{i2}^{(U,U)}$ $\forall i$ since $\frac{\partial f_2(\cdot)}{\partial S_{i2}^{(T,T)}} \leq \frac{\partial f_2(\cdot)}{S_{i2}^{(U,T)}}$, indicating that $\tau_2^{(T,T)} \leq \tau_2^{(U,T)}$; a contradiction. ■

Proposition 4 *If ability is additively separable, then school inputs are independent of student ability.*

Proof: If ability, α_i, is additively separable, then $h[A_{it} - A_{i,t-k} - \zeta_k] = h[f_t(\alpha_i, S_{i1}, ..., S_{it}) - f_t(\alpha_i, S_{i1}, ..., S_{i,t-k}) - \zeta_k] = h[\alpha_i + f_t(S_{i1}, ..., S_{it}) - \alpha_i - f_t(S_{i1}, ..., S_{i,t-k}) - \zeta_k] = h[f_t(S_{i1}, ..., S_{it}) - f_t(S_{i1}, ..., S_{i,t-k}) - \zeta_k]$ $\forall t, k$. The first-order conditions are therefore independent of α_i. ■

Proposition 5 *If the test score improvement target $\zeta_k = \zeta$ $\forall k$, student achievement is (weakly) increasing in the number of years (k) that the test score used to determine the student-specific targets is lagged.*

Proof: The FOC in the k-type scheme $\forall t \leq k$ is:

$$h[A_{it} - A_{i0} - \zeta]\frac{\partial A_{it}}{S_{it}} + ... + h[A_{ik} - A_{i0} - \zeta]\frac{\partial A_{ik}}{S_{it}} \geq c'(S_{it})$$

while for a $k-1$-type scheme is:

$$h[A_{it} - A_{i0} - \zeta]\frac{\partial A_{it}}{S_{it}} + ... + h[A_{i,k-1} - A_{i0} - \zeta]\frac{\partial A_{i,k-1}}{S_{it}} + h[A_{ik} - A_{i1} - \zeta]\left(\frac{\partial A_{ik}}{S_{i1}} - \frac{\partial A_{i1}}{S_{i1}}\right) \geq c'(S_{i1}) \quad \text{for } t = 1,$$
$$h[A_{it} - A_{i0} - \zeta]\frac{\partial A_{it}}{S_{it}} + ... + h[A_{ik} - A_{i0} - \zeta]\frac{\partial A_{ik}}{S_{it}} \geq c'(S_{it}) \quad \text{for } 1 < t \leq k.$$

The dynamic distortion for $t = 1$ in the $k-1$-type scheme (from the $-\frac{\partial A_{i1}}{S_{i1}}$ term) gives us that S_{i1} is greater under the k-type scheme than the $k-1$-type scheme. Since the FOCs for periods $1 < t \leq k$ are identical (a similar argument holds for $t > k$), achievement in period $t \leq k$ will be higher under the k-type scheme. ■

B Identifying Teacher Effort

The estimation procedure builds upon the approach in Macartney et al. (2016) and consists of three steps:

1. Teacher Fixed-Effects: As a first step, teacher-year fixed-effects are calculated by regressing contemporaneous test scores on prior test scores and other student characteristics. Formally, I calculate teacher-year fixed effects separately for each grade using the regression:

$$y_{ijt} = \omega(y_{ij,t-1}) + \beta Z_{ijt} + q_{jt} + v_{ijt}\,, \tag{B.1}$$

where y_{ijt} are test scores of student i in class j at time t, $\omega(\cdot)$ is some flexible functional form, q_{jt} is teacher quality, Z_{ijt} is a control vector that includes student race, gender, disability status, gifted status and limited English-proficiency status, and v_{ijt} is the error term. Teacher-year fixed effects are given by:

$$\hat{q}_{jt} = \sum_{i=1} \frac{\hat{y}_{ijt} - \hat{\omega}(y_{ij,t-1}) - \hat{\beta} Z_{ijt}}{n_{jt}} = a_j + e_{jt}(g_{jt}) + \bar{v}_{jt}\,, \tag{B.2}$$

where the second equality comes from the fact that teacher year fixed effects consist of ability and incentive-invariant effort, a_j, and incentive varying effort, which I label as e_{jt}. Teacher effort in this setting is a function the proportion of subgroup-specific students, g_{jt}, in teacher j's class at time t.

2. Estimating Teacher Ability: Consistent with much of the literature, teacher ability is assumed to be fixed over time. Each teacher's ability is then calculated by averaging teacher-year fixed-effects over all non-current years. Formally, for a teacher with T_j years of teaching, her ability at time τ is:

$$a_{j\tau} = \sum_{\tau \neq t}^{T_j} \frac{\hat{q}_{j\tau}}{T_j - 1} - \sum_{\tau \neq t}^{T_j} \frac{e_{jt}(g_{jt})}{T_j - 1} - \bar{v}_{jt} = \sum_{\tau \neq t}^{T_j} \frac{\hat{q}_{j\tau}}{T_j - 1} + \bar{\eta}_{jt}\,, \tag{B.3}$$

where the term $\bar{\eta}_{jt}$ captures the error term and average effort over non-current years.

3. Identifying Subgroup-Specific Effort: To identify the increase in teacher effort associated with facing subgroup-specific accountability, I approximate the underlying effort function with a linear functional form so that average subgroup-specific effort is given by $\tilde{e} = \pi_2 g_{jt}$. I now use the RD design to capture accountability-induced effort responses. First, define the teacher-year fixed-effect residual, ζ_{jt}, as:

$$\zeta_{jt} = \hat{q}_{jt} - a_{j\tau} + \bar{\eta}_{jt} \tag{B.4}$$

To implement the design, I define the estimated teacher-year fixed effect residual $\hat{\zeta}_{jt} = \hat{q}_{jt} - \hat{a}_{j\tau}$.[67] Then, I obtain teacher effort as a function of the number of students in a given subgroup in her class, $\hat{\zeta}_{jt}$. From this, I can take advantage of the fact that teachers in a school with fewer than forty students face no accountability under NCLB for those students, while teachers in a school with forty or more students face accountability pressure for those students. I therefore contrast the residual for teachers in a school with more than forty students in a subgroup, ζ_{jt}^{RHS}, with the residual for teachers where there are fewer than forty students in that subgroup in the school, ζ_{jt}^{LHS}. Since subgroup accountability only holds for schools with more than forty students in a subgroup, we have:

$$\zeta_{jt}^{LHS} = \hat{q}_{jt} - a_j + \bar{\eta}_{jt}, \tag{B.5}$$

$$\zeta_{jt}^{RHS} = \hat{q}_{jt} - a_j + \bar{\eta}_{jt} + \pi_2 g_{jt}\,. \tag{B.6}$$

Subgroup-specific effort is then defined as the difference between these two relationships, on either side of the forty student threshold:

$$\tilde{e}_{jt} = \pi_2 g_{jt} = \zeta_{jt}^{RHS} - \bar{\eta}_{jt}^{RHS} - \zeta_{jt}^{LHS} + \bar{\eta}_{jt}^{LHS}\,. \tag{B.7}$$

The forty or more line, the less than forty line, and the difference between lines are reported in Appendix Figure A7, giving the estimated $\hat{\pi}_2$ coefficients for Equations B.5, B.6, and B.7, respectively.

[67] I also control for the number of 'marginal' students in the classroom to capture any increase in teacher effort caused by students with different ability levels being assigned in a systematic way to their class.

C Literature Exploring Dynamic Complementarities

Paper	Data	Research Design	Findings
Todd and Wolpin (2007)	Use the NLSY/79-CS, a sample of all children born to individuals in the NLSY/79. The data provide test scores, measures of the child's home environment, mother characteristics, and county-level measures of school quality.	Estimate a cumulative production function for children's cognitive achievement that allows achievement to depend on innate ability, mother's abilities, and the lifetime history of family and school inputs. Identify the model with value-added specifications and child and sibling fixed-effects.	Estimates strongly support the notion that skill acquisition is a cumulative process. School input effects are imprecisely measured. Specification tests point toward incorporating lagged inputs, consistent with there being input complementarities across time.
Cunha and Heckman (2008)	Estimate the technology on a sample of 2,207 first-born white children from the NLSY/79. Use multiple measures of cognitive and non-cognitive skills as outputs (and inputs) in their estimation.	Estimate a dynamic factor model that exploits cross-equation restrictions to secure identification. To deal with the endogeneity of inputs, use next-period skills and investments as an instrument for current period skills and investments.	Parental inputs have different effects at different stages of the child's life cycle, with cognitive skills affected more at early ages and noncognitive skills affected more at later ages.
Heckman, Moon, Pinto, Savelyev, and Yavitz (2010)	Use data from the HighScope Perry Preschool Program, which was conducted in the 1960s and targeted disadvantaged children in Ypsilanti, Michigan.	Use statistical corrections to the compromised HighScope Perry Preschool Program randomization to find the effects of early investments on later outcomes.	Perry Preschool treatment effects are substantial. Find the largest effects on cognitive achievement among those at the top of the distribution, consistent with a complementarity between early human capital and later investments.
Cunha, Heckman, and Schennach (2010)	Estimate the technology on a sample of 2,207 first-born white children from the NLSY/79. Use multiple measures of cognitive and non-cognitive skills as outputs (and inputs) in their estimation.	Paper formulates and estimates multistage production functions for children's skills using a dynamic factor model. To deal with the endogeneity of inputs, use lagged parental income as an instrument for current period child investment.	Find that substitutability decreases in later stages of the life cycle. It is therefore optimal to invest relatively more in the early stages of childhood than in later stages. Find that the technology features parental input complementarity across time.

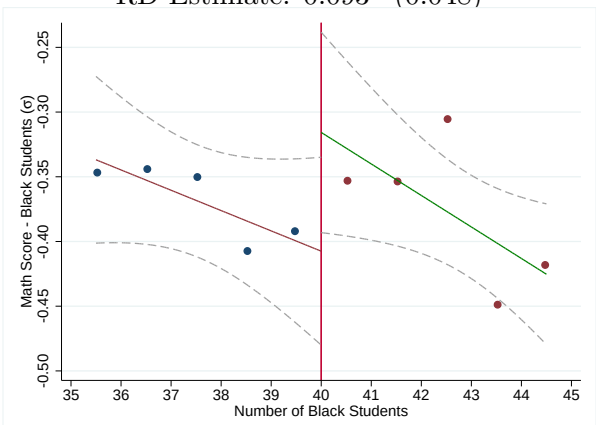
Aizer and Cunha (2012)	Use data from the National Collaborative Perinatal Project, which contain comprehensive information for roughly 59,000 births between 1959 and 1965. Use the Bayley scores of development as a measure of a child's initial human capital stock and IQ or test scores as later measures of human capital.	Exploit exogenous variation in investment from the roll-out of Head Start in 1966 across children in the same household. Then search for differential effects among children with different stocks of human capital.	Find that preschool enrollment has a positive and significant impact on four-year IQ for all, but that the impact is largest for those with higher early stocks of cognitive human capital.
Lubotsky and Kaestner (2016)	Use data from the Early Childhood Longitudinal Study Kindergarten Class of 1998-99, which cover approximately 15,000 students from a random sample of over 1,000 kindergarten classrooms. Use test scores as cognitive skill measures and also have five measures of noncognitive skills. They also use the NSLY/79.	Instrument the actual kindergarten entrance age with the entrance age of the child if she had followed the state entrance law. Then investigate whether children who start kindergarten at an older age – thus having a higher stock of skill – have faster test score growth than children starting kindergarten at a younger age.	Children beginning kindergarten at an older age have higher measures of cognitive and noncognitive achievement at the start of kindergarten and have their cognitive scores grow at a faster rate (by 0.11σ) during kindergarten and first grade – consistent with complementarities in human capital accumulation.
Malamud, Pop-Eleches, and Urquiola (2016)	Use administrative data from Romania containing all the children who were allocated to a high school in the years 2005 and 2006.	Combine a shock to initial child ability with a later shock to schooling inputs. The shock to initial ability comes from a difference-in-differences design taking advantage of a change in the abortion law, while the schooling inputs shock comes from a regression discontinuity design where school quality was effectively randomized.	Find no direct evidence of dynamic complementarities, noting that the authors also discover suggestive evidence that parents and children behave in ways that would tend to undo dynamic complementarities.
Johnson and Jackson (2017)	Use individual level data from the Panel Study of Income Dynamics which consist of 13,381 individuals born between 1950 and 1976.	Combine policy-induced changes in pre-school (Head Start) spending and school-finance-reform-induced changes in public K-12 school spending during childhood as their two exogenous shocks to investigate dynamic complementarities.	Look at high school graduation, years of schooling, incarceration and adult earnings as outcomes. Find evidence of dynamic complementarities in that the marginal effect of a 10 percent increase in K-12 school spending are 50-100 percent larger when Head Start spending is at the 75th percentile rather than the 25th percentile.

D Appendix Figures and Tables

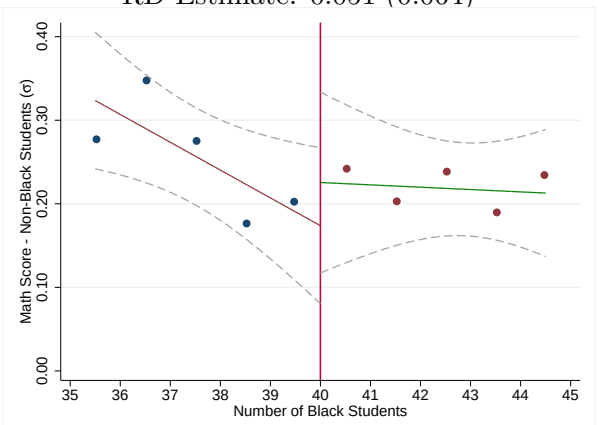
Figure A1: Reduced-Form (Math)

Black Subgroup: Diff-in-Disc Estimate: 0.042 (0.053)

(a) Student is Black
RD Estimate: 0.093* (0.048)

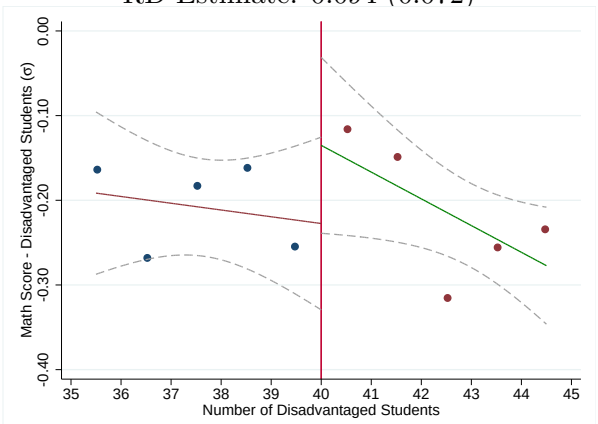


(b) Student is Not Black
RD Estimate: 0.051 (0.064)

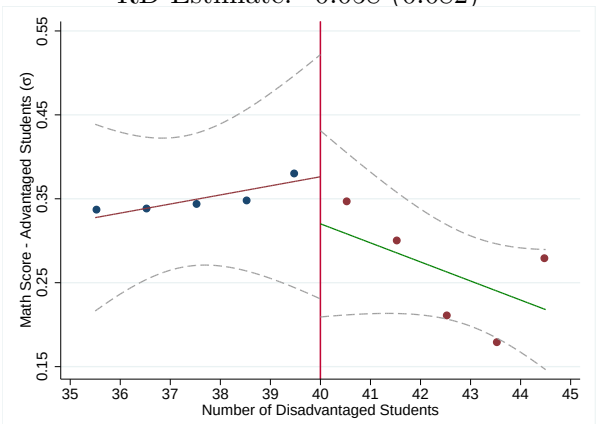


Disadvantaged Subgroup Diff-in-Disc Estimate: 0.152* (0.082)

(c) Student is Disadvantaged
RD Estimate: 0.094 (0.072)

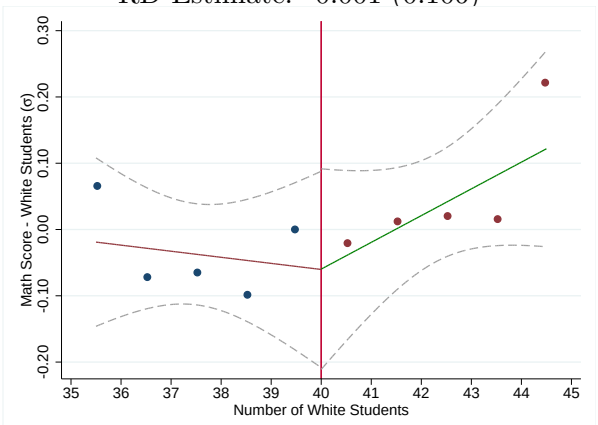


(d) Student is Not Disadvantaged
RD Estimate: -0.058 (0.082)

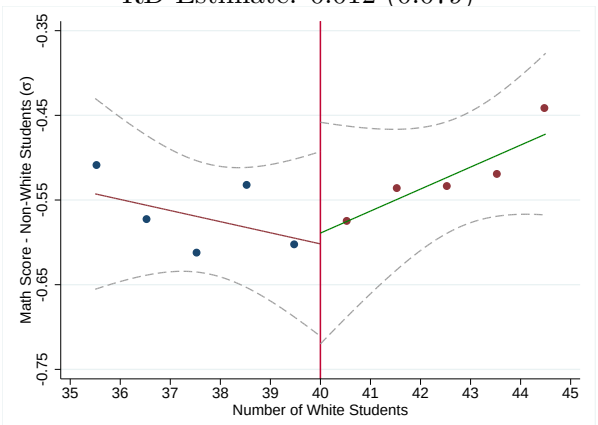


White Subgroup Diff-in-Disc Estimate: -0.013 (0.072)

(e) Student is White
RD Estimate: -0.001 (0.106)



(f) Student is Not White
RD Estimate: 0.012 (0.079)

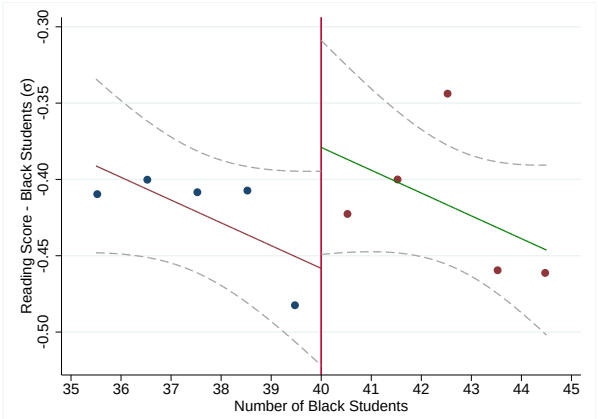


Notes: Figures for the black, disadvantaged and white subgroup are based on 558, 330 and 248 observations, respectively. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 90% confidence intervals, with standard errors clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

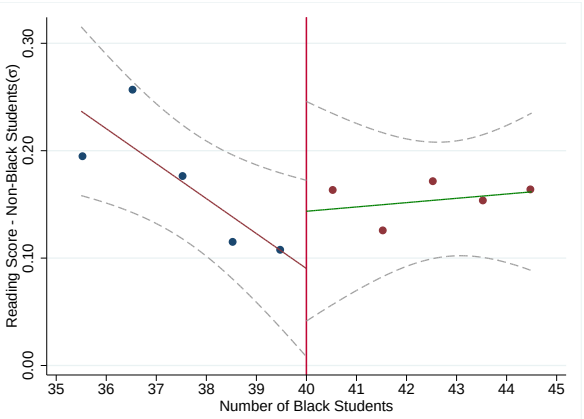
Figure A2: Reduced-Form (Reading)

Black Subgroup: Diff-in-Disc Estimate: 0.028 (0.051)

(a) Student is Black
RD Estimate: 0.080* (0.044)

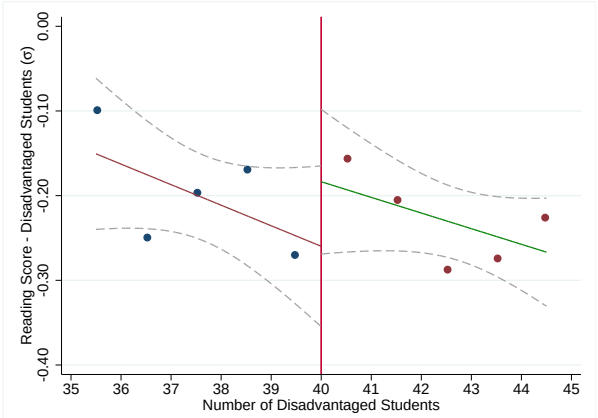


(b) Student is Not Black
RD Estimate: 0.053 (0.060)

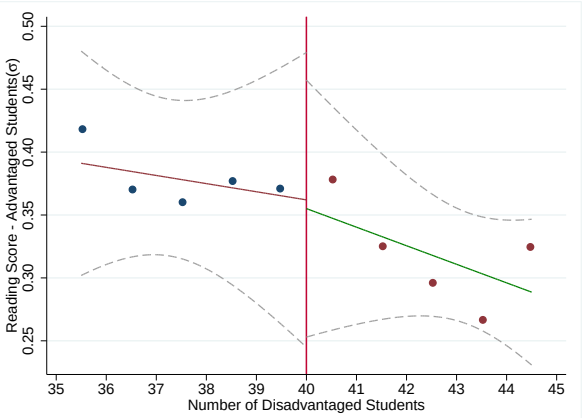


Disadvantaged Subgroup Diff-in-Disc Estimate: 0.086 (0.081)

(c) Student is Disadvantaged
RD Estimate: 0.078 (0.064)

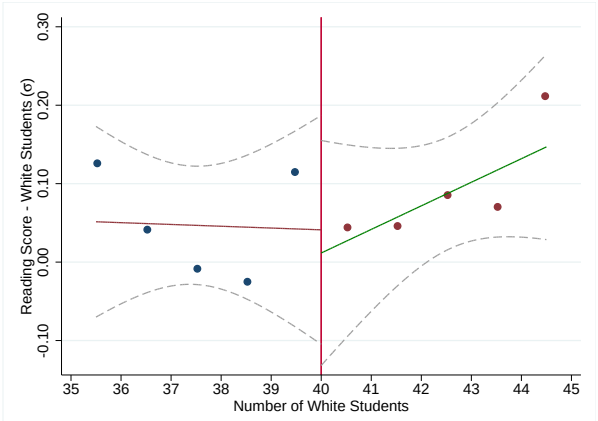


(d) Student is Not Disadvantaged
RD Estimate: -0.007 (0.072)

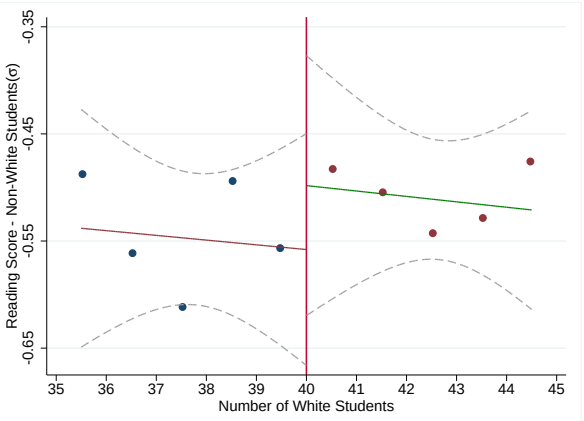


White Subgroup Diff-in-Disc Estimate: -0.093 (0.072)

(e) Student is White
RD Estimate: -0.032 (0.103)



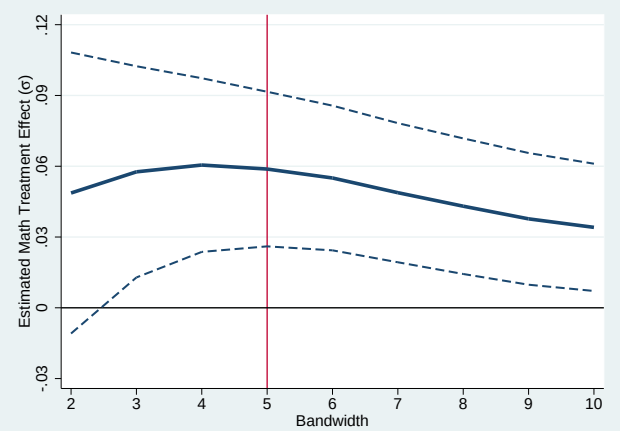
(f) Student is Not White
RD Estimate: 0.061 (0.074)



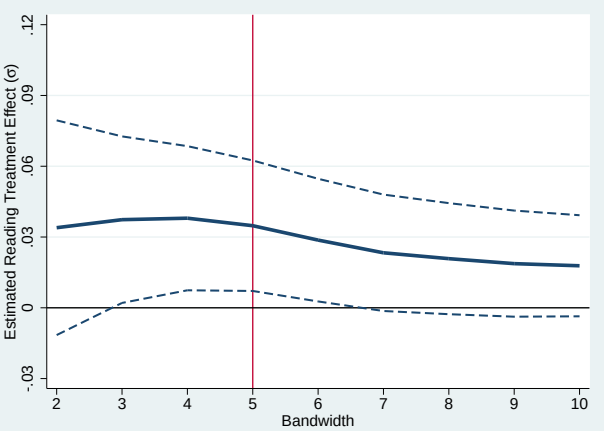
Notes: Figures for the black, disadvantaged and white subgroup are based on 558, 330 and 248 observations, respectively. Each RD estimate is from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Dashed lines represent 90% confidence intervals, with standard errors clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A3: Bandwidth Sensitivity

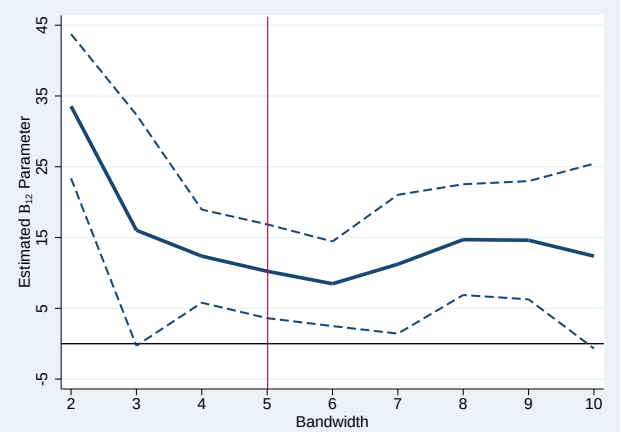
(a) Math Scores (σ)



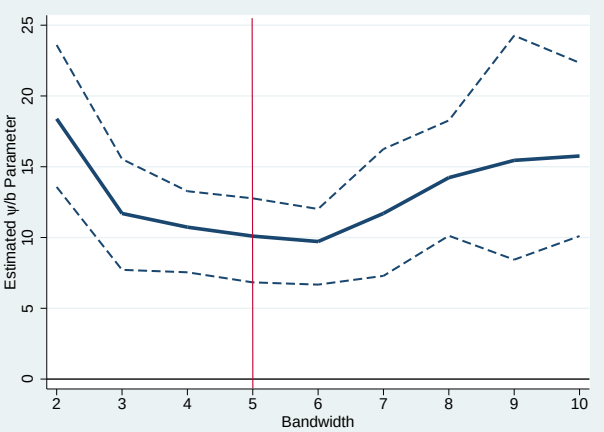
(b) Reading Scores (σ)



(c) Structural Parameter Estimate: $\hat{\beta}_{12}$



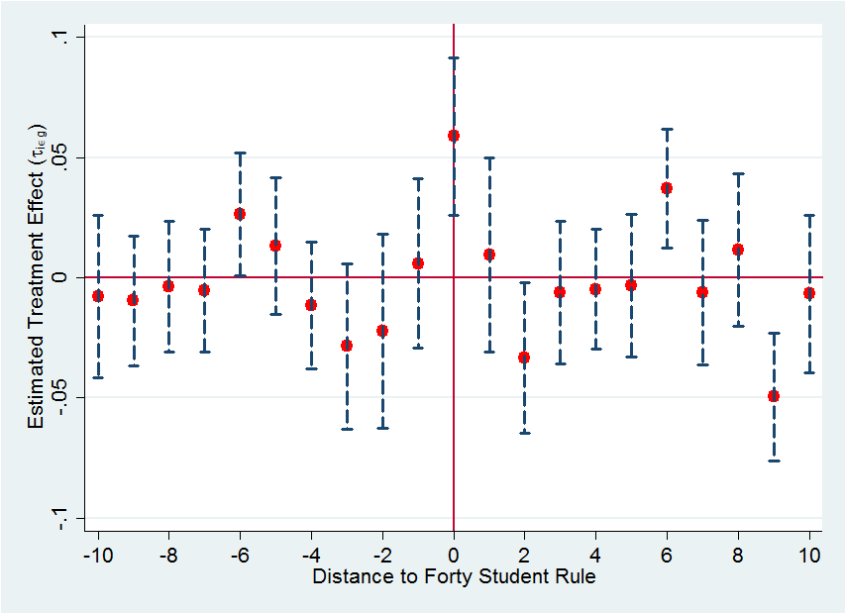
(d) Structural Parameter Estimate: $\frac{\hat{\psi}}{b}$



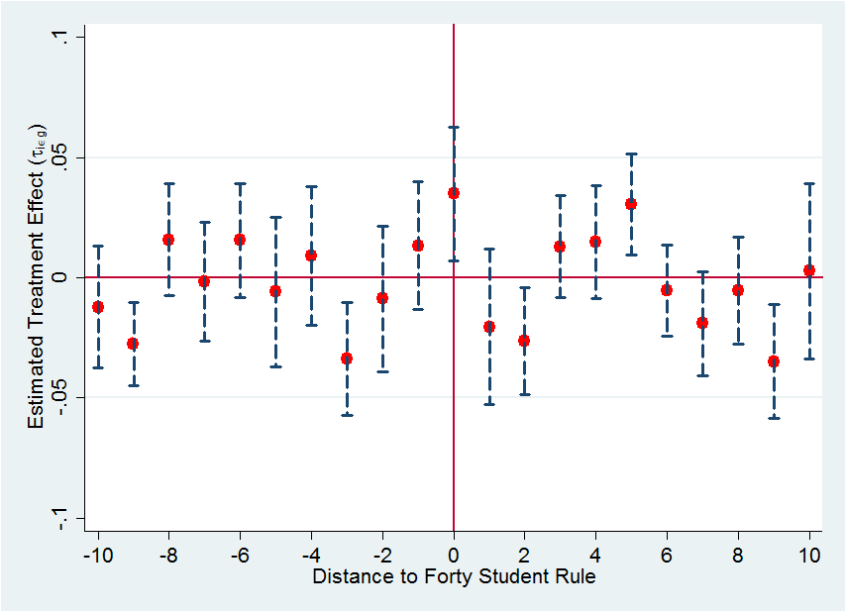
Notes: Figures A3(a) and A3(b) plot RD estimates for $\tau_{i\in g}$ from Table 2 for a variety of bandwidth values. Covariates are included. Similarly, Figures A3(c) and A3(d) plot structural parameter estimates for a variety of bandwidth values. The bandwidth of one is omitted due to high standard errors. The vertical line represents the chosen bandwidth and the horizontal line represents an estimate of zero. The dashed lines are 90% confidence intervals. Following Lee and Card (2008), standard errors are clustered on sixty student-by-subgroup clusters for Figures A3(a) and A3(b) and are bootstrapped for Figures A3(c) and A3(d).

Figure A4: Placebo Subgroup Rules

(a) Math Scores (σ)



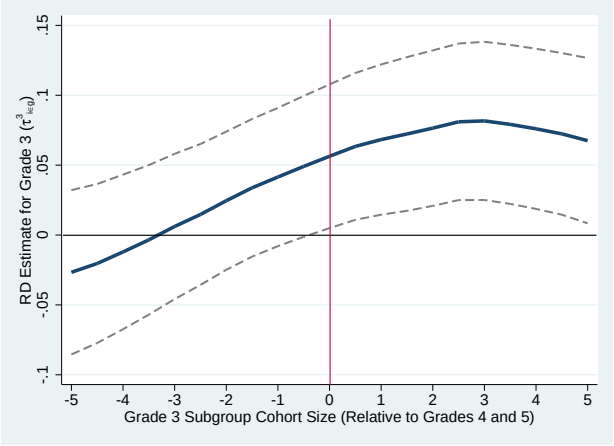
(b) Reading Scores (σ)



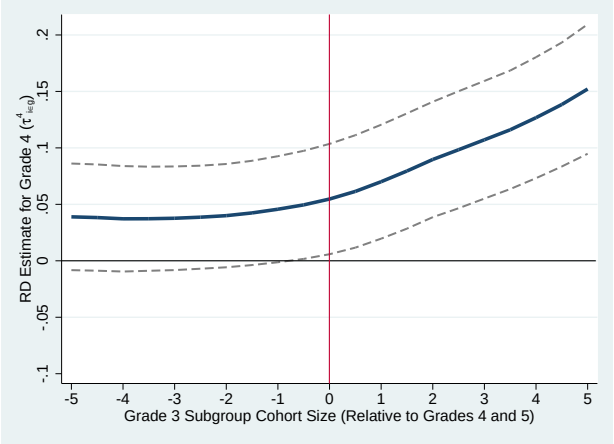
Notes: The figures plot RD point estimates for $\tau_{i\in g}$ from Table 2 for a variety of placebo subgroup rules. The placebo subgroup rules vary from thirty to fifty students. Covariates are included. The bandwidth used is five and the vertical line represents the true subgroup rule of forty. The horizontal line represents a parameter estimate of zero and the vertical dashed lines are 90 percent confidence intervals. Standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008).

Figure A5: Estimated Treatment Effects by Grade and Expectations

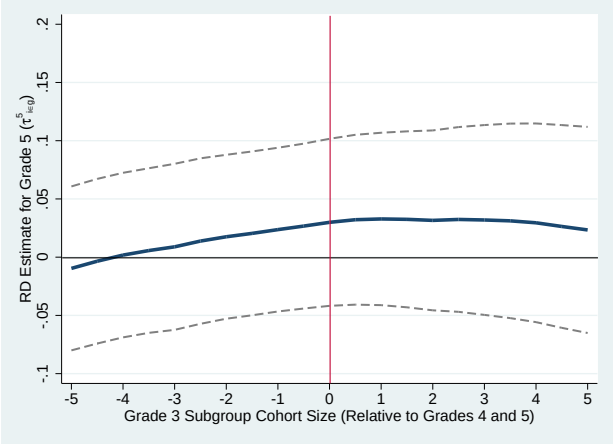
(a) Grade 3 ($\hat{\tau}^3_{i\in g}$)



(b) Grade 4 ($\hat{\tau}^4_{i\in g}$))



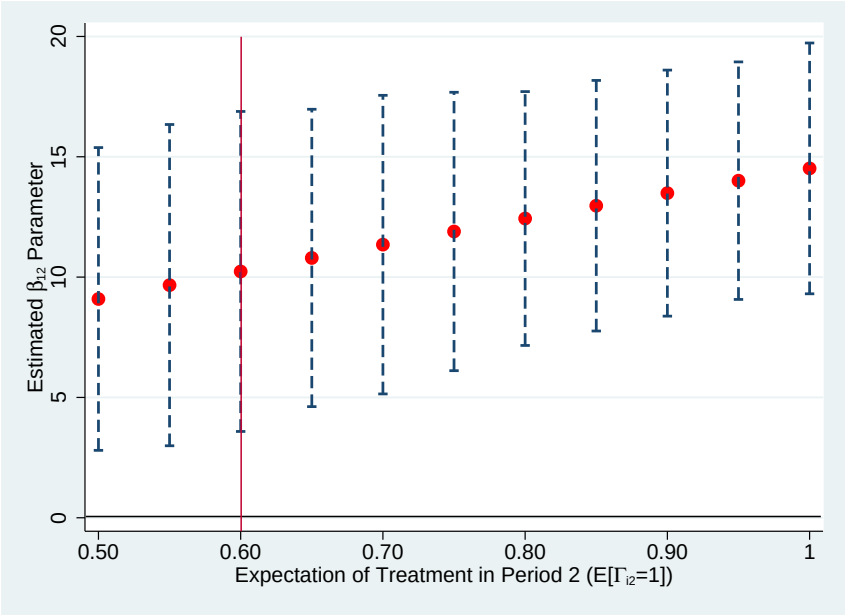
(c) Grade 5 ($\hat{\tau}^5_{i\in g}$)



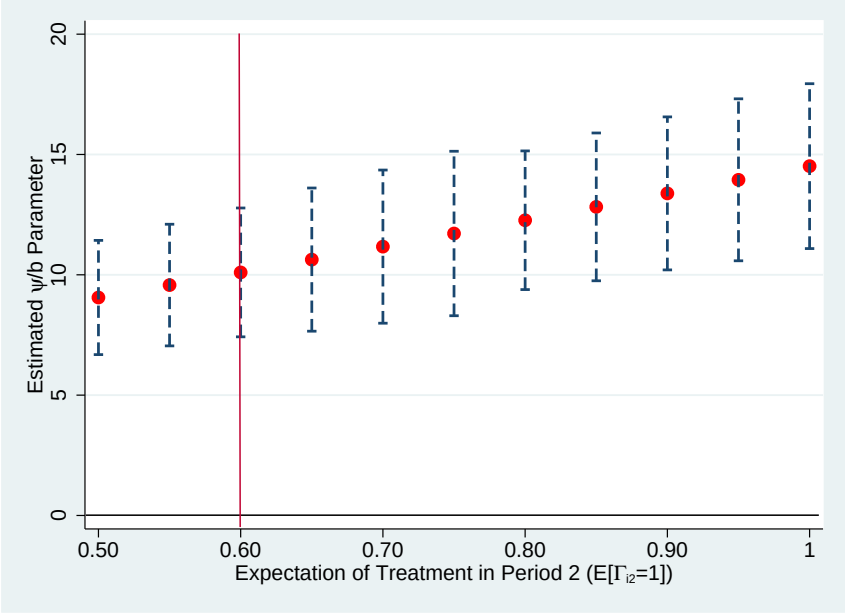
Notes: The figures plot $\hat{\tau}^3_{i\in g}$, $\hat{\tau}^4_{i\in g}$, and $\hat{\tau}^5_{i\in g}$ for various grade 3 subgroup cohort sizes (relative to grade 4 and grade 5), which schools should use to form their expectations of future treatment. The x-axis represents the number of students belonging to a subgroup in grade 3, divided by the number of students in that subgroup in grades 3, 4, and 5. The unit for the x-axis is percentage points. The vertical line denotes zero on the x-axis, which occurs when the number of grade 3 students in the subgroup is identical to the average subgroup cohort size in grades 4 and 5 (i.e. thirty-three percent). The horizontal line denotes a RD estimate of zero. Dashed lines represent 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008).

Figure A6: Structural Estimates under Uncertainty

(a) Structural Parameter Estimate: $\hat{\beta}_{12}$



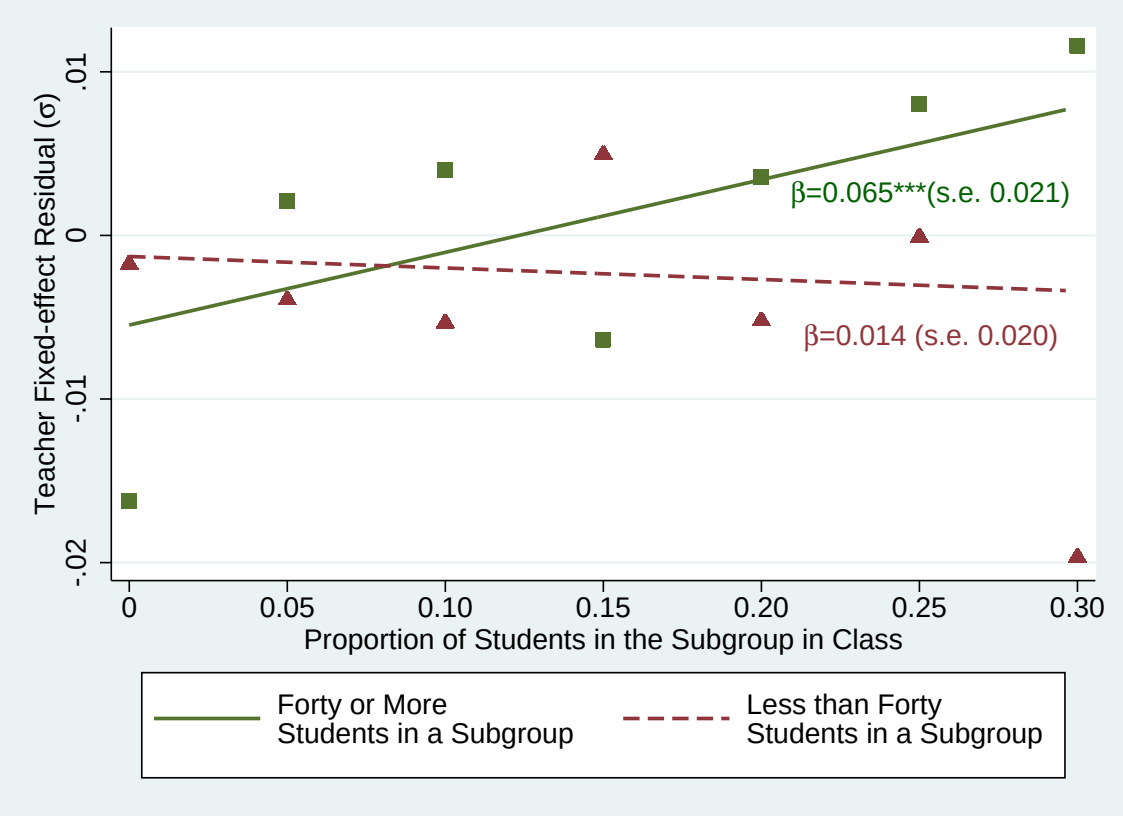
(b) Structural Parameter Estimate: $\frac{\hat{\psi}}{b}$



Notes: The panels plot the structural estimates of $\hat{\beta}_{12}$ and $\frac{\hat{\psi}}{b}$ for a variety of school expectations for the probability of period two treatment. See Section 8.1 for a further description. The bandwidth used is five and the vertical line represents the rational expectation case used in the structural estimation (which corresponds to the parameter estimates in Column (1) of Appendix Table A4). The horizontal line represents a parameter estimate of zero and the vertical dashed lines are 90 percent confidence intervals. Standard errors are bootstrapped.

Figure A7: Identifying Effort

(a) Effort
Difference between lines: 0.051* (s.e. 0.030)

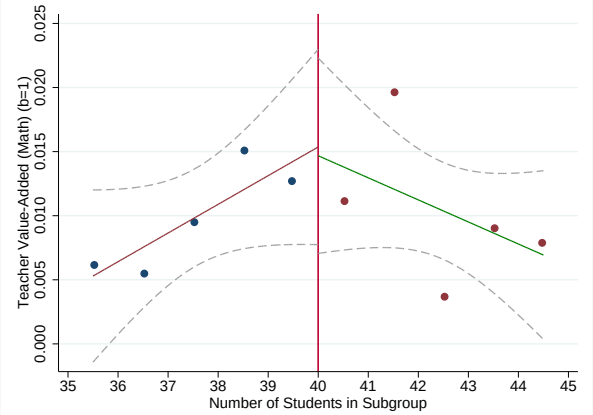


Notes: Squares and triangles represent binned means for the forty or more and less than forty students in a subgroup lines, respectively. Figure A7 is based on 3,861 and 4,391 classroom-year observations for the forty or more and less than forty lines, respectively. Due to few observation after thirty percent, the x-axis is restricted to below thirty percent. Slope estimates are indicated below each line and are for the full sample. Standard errors are clustered at the school level due to an insufficient number of student-by-subgroup clusters. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

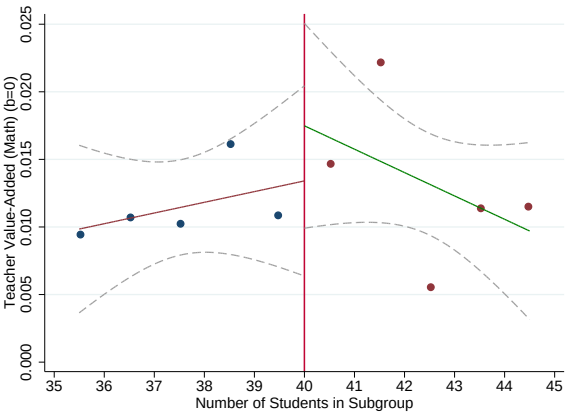
Figure A8: Alternative Mechanisms

Teacher Value-Added

(a) Student belongs to Subgroup
RD Estimate: -0.000 (0.006)

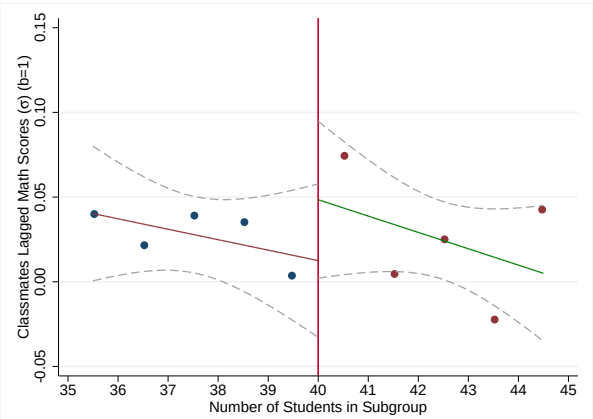


(b) Student does not belong to Subgroup
RD Estimate: 0.004 (0.005)

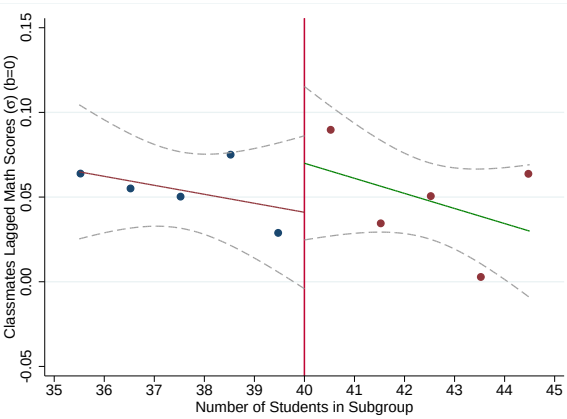


Classmates

(c) Student belongs to Subgroup
RD Estimate: 0.037 (0.029)

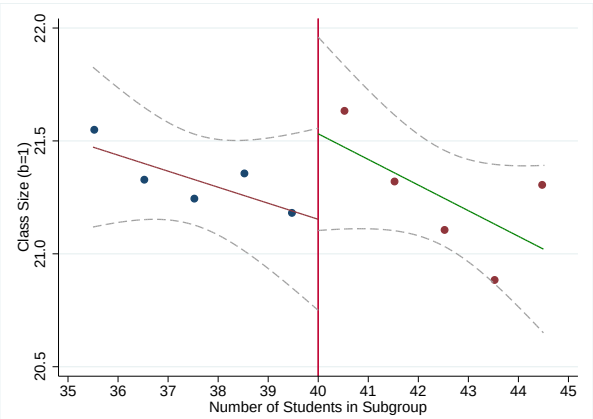


(d) Student does not belong to Subgroup
RD Estimate: 0.030 (0.031)

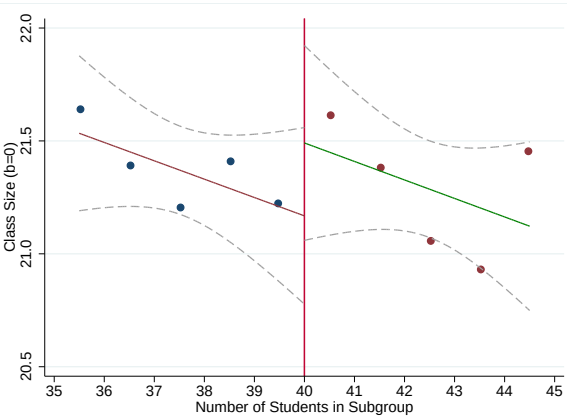


Class Size

(e) Student belongs to Subgroup
RD Estimate: 0.347 (0.216)



(f) Student does not belong to Subgroup
RD Estimate: 0.330 (0.211)



Notes: All figures are based on 1,626 observations. Figures and RD estimates control for subgroup fixed effects. 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 90% confidence intervals with standard errors clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A1: Tests of Discontinuities in Observable Covariates

	Grade 2 Math Scores (1)	Grade 2 Reading Scores (2)	Percent Black (3)	Percent White (4)	Percent Hispanic (5)	Percent Asian (6)	Percent Disad-vantaged (7)	Percent EL[1] (8)	Percent SWD[2] (9)	Percent Gifted (10)	Percent Male (11)
A. Students Belong to Subgroup											
# in subgroup≥40	0.030	0.048*	-2.30	2.30	0.57	0.03	-0.09	0.29	-0.23	-0.91	-0.03
	(0.030)	(0.025)	(1.61)	(1.67)	(0.52)	(0.20)	(1.42)	(0.98)	(0.57)	(0.79)	(0.81)
Joint Significance Test (Prob>χ^2)			0.195								
B. Students do Not Belong to Subgroup											
# in subgroup≥40	0.048	0.047*	-2.17*	2.14	-0.72	0.25	-2.99*	-0.59	-0.28	0.21	-0.55
	(0.032)	(0.027)	(1.11)	(1.52)	(0.86)	(0.39)	(1.59)	(0.62)	(0.28)	(1.03)	(0.50)
Joint Significance Test (Prob>χ^2)			0.893								
C. Difference in Covariates											
# in subgroup≥40	-0.018	0.000	-0.13	0.17	1.29	-0.23	2.90*	0.88	0.05	-1.11	0.52
	(0.034)	(0.022)	(1.41)	(1.46)	(1.00)	(0.42)	(1.57)	(1.17)	(0.63)	(0.78)	(0.98)
Observations	3,292	3,292	3,292	3,292	3,292	3,292	3,292	3,292	3,292	3,292	3,292
Joint Significance Test (Prob>χ^2)			0.782								

[1] EL is an acronym for 'English learners.'
[2] SWD is an acronym for 'students with learning disabilities.'

Notes: The number of observations is given in Panel C: Panels A and B have half that number of observations each. Each cell represents results for a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Subgroup fixed-effects are included. Standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). The joint significance test reports the p-value from a hypothesis test with a null hypothesis that all covariates are jointly continuous at the threshold for that panel. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A2: Tests of Discontinuities and Level Differences in Observable Covariates, by Treatment Status for Period Two

	Percent Black (1)	Percent White (2)	Percent Hispanic (3)	Percent Asian (4)	Percent Disad-vantaged (5)	Percent EL[1] (6)	Percent SWD[2] (7)	Percent Gifted (8)	Percent Male (9)
A. Treated Last Year									
# in subgroup≥40	0.77	-2.99	2.38	-0.49	0.14	-1.44	1.12	-1.49	1.18
	(1.55)	(3.60)	(2.38)	(0.31)	(4.04)	(1.81)	(0.97)	(0.97)	(2.08)
Joint Significance Test (Prob>χ^2)			0.513						
B. Untreated Last Year									
# in subgroup≥40	-0.25	-0.65	-0.43	1.13	2.27	-2.97*	1.43*	2.38	1.57
	(1.09)	(1.81)	(0.40)	(1.13)	(1.93)	(1.70)	(0.77)	(1.72)	(1.74)
Joint Significance Test (Prob>χ^2)			0.122						
C. Difference in Covariates									
# in subgroup≥40	1.02	-2.34	2.81	-1.62	-2.13	1.53	-0.31	-3.86**	-0.39
	(1.29)	(2.51)	(2.54)	(1.42)	(3.90)	(1.52)	(1.09)	(1.82)	(1.87)
Joint Significance Test (Prob>χ^2)			0.083						
D. Difference in Levels									
Treated Last Year	0.64	-0.56	0.47	-0.48	2.69	0.94	-0.71	-0.55	-0.32
	(0.40)	(0.58)	(0.63)	(0.38)	(1.64)	(0.94)	(0.48)	(0.81)	(0.83)
Observations	1,127	1,127	1,127	1,127	1,127	1,127	1,127	1,127	
Joint Significance Test (Prob>χ^2)			0.281						

[1] EL is an acronym for 'English learners.'
[2] SWD is an acronym for 'students with learning disabilities.'

Notes: This table looks for discontinuities or a difference in levels between period two (i.e. grade 4) students in schools that were treated and were not treated in the prior year. The number of observations is given in Panel D: Panel C has the same number of observations, while Panels A and B have about half that number of observations each. Panels A and B check for discontinuities in covariates at the threshold for students who were and were not treated in the prior period. Panel C then tests whether the difference-in-discontinuities between Panels A and B are statistically different. Finally, Panel D checks whether the samples underlying Panel A and B differ significantly. For Panels A, B, and C, each cell represents results for a separate local linear regression allowing for different functions on either side of the threshold, while each cell in Panel D represents a difference-in-means test. The bandwidth used is five. Subgroup fixed effects are included. Standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). The joint significance test reports the p-value from a hypothesis test with a null hypothesis that all coefficients are jointly zero for that panel. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A3: Test Scores by Subgroup

	Black	White	Hispanic	Asian	Multi-Racial	Native American	Disadvantaged
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean (S.D.) Test Scores							
Math Scores (σ)	-0.473	0.318	-0.206	0.568	0.019	-0.297	-0.349
	(0.257)	(0.335)	(0.311)	(0.612)	(0.456)	(0.450)	(0.260)
Reading Scores (σ)	-0.425	0.308	-0.356	0.281	0.053	-0.309	-0.359
	(0.241)	(0.297)	(0.322)	(0.584)	(0.437)	(0.450)	(0.233)
Observations (student-level)	418,659	844,606	117,004	31,010	45,150	19,052	701,183

Notes: The entire sample is used. The table reports average test scores for students in each subgroup across all grades.

Table A4: Robustness: Estimates Using
Different Functional Forms

Outcome Variable: Math Test Scores (σ)

	Linear (1)	Quadratic (2)	Triangular Kernel (3)
A. Reduced-Form Estimates			
$\tau_{i \in g}$	0.053**	0.069**	0.059***
	(0.021)	(0.031)	(0.020)
$\tau_{i \notin g}$	0.008	0.036	0.017
	(0.018)	(0.028)	(0.017)
τ_{diff}	0.046***	0.033	0.042***
	(0.016)	(0.024)	(0.015)
B. Structural Estimates			
β_{12}	10.24***	12.81***	7.46**
	(3.22)	(2.93)	(3.79)
$\frac{\psi}{b}$	10.10***	10.67***	9.00***
	(1.55)	(1.10)	(1.65)

Notes: Results for the reduced form estimates under a triangular kernel functional form are identical to the results reported in column (2) of Table 2. Covariates are included. The bandwidth used is five. For Panel A, standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). For Panel B, standard errors are bootstrapped with 250 repetitions. Significance levels: *** 1 percent; ** 5 percent; * 10 percent.

Table A5: Regression-Discontinuity Estimates of Subgroup-Specific Accountability on Student Achievement by Grade

Outcome: Math Scores (σ)	Grade 3 (1)	Grade 4 (2)	Grade 5 (3)
Panel A. Students in Subgroup			
$\tau_{i \in g}$	0.048**	0.073***	-0.002
	(0.024)	(0.024)	(0.025)
Panel B. Students Not in Subgroup			
$\tau_{i \notin g}$	-0.017	0.028	-0.020
	(0.022)	(0.018)	(0.019)
Panel C. Difference-in-Discontinuity			
τ_{diff}	0.066***	0.045**	0.018
	(0.020)	(0.018)	(0.020)
Observations	1,798	2,317	2,126

Notes: The number of observations is given in Panel C: Panel A and B have half that number of observations each. The 2002-03 school year is omitted to maintain the same sample in Table 3. Grade 3 results are missing for the 2005-06 school year due to missing data on the math pre-test for that year. The bandwidth used is five. Covariates include controls for prior test scores, gender, ethnicity, gifted status, English learner status, free lunch status, and grade configuration, subgroup and year fixed effects. Standard errors are clustered on sixty student-by-subgroup clusters, following Lee and Card (2008). ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A6: Estimates of Within-Classroom Spillovers

	Outcome: Math Scores (σ)		
	RHS (i.e. $X_{sgt} \geq 40$)	LHS (i.e. $X_{sgt} < 40$)	Difference
	(1)	(2)	(3)
Panel A. Students not belonging to Subgroup			
Proportion of Accountable Students in a Class	-0.022	0.052	-0.074
	(0.046)	(0.042)	(0.059)
Panel B. Students belonging to Subgroup			
Proportion of Accountable Students in a Class	0.020	-0.023	0.043
	(0.058)	(0.055)	(0.074)
Observations (school-subgroup-class-year)	5,271	5,733	11,004

Notes: RHS and LHS are acronyms for right- and left-hand side of the forty student threshold. Column (1) reports the relationship between the number of students in a subgroup in a class when there are forty or more of these students at the school. Column (2) reports the same relationship when there are less than forty of these students at the school. Sample is restricted to the RD sample and a triangular kernel is used to put more weight on observations closer to forty students in the subgroup at the school. Covariates are included. Standard errors are clustered at the school level due to an insufficient number of student-by-subgroup clusters. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.