SPENDING MORE OF THE SCHOOL DAY IN MATH CLASS: EVIDENCE FROM A REGRESSION DISCONTINUITY IN MIDDLE SCHOOL

Eric Taylor
Stanford University

520 Galvez Mall
CERAS Building, 509
Stanford, CA 94305
erictaylor@stanford.edu
617-595-8191

June 3, 2014

Abstract: For students whose math skills lag expectations, public schools often increase the fraction of the school day spent on math instruction. Studying middle-school students and using regression discontinuity methods, I estimate the causal effect of requiring two math classes—one remedial, one regular—instead of just one class. Math achievement grows much faster under the requirement, 0.16-0.18 student standard deviations. Yet, one year after returning to a regular one-class schedule, the initial gains decay by as much as half, and two years later just one-third of the initial treatment effect remains. This pattern of decaying effects over time mirrors other educational interventions—assignment to a more skilled teacher, reducing class size, retaining students—but spending more time on math carries different costs. One cost is notable, more time in math crowds out instruction in other subjects.

Key words: education production, remedial education, math achievement, effect persistence

Many students' math skills fall far short of expectations. In the most recent National Assessment of Educational Progress (NAEP), two-thirds of 8th grade (14-15 year old) students were not "Proficient" in math; more concretely, they likely could not, for example, draw lines of symmetry in a geometry problem or solve a measurement problem involving unit conversions (National Center for Education Statistics 2011). In response, many public schools have begun increasing the quantity of time that struggling students spend in math class, with the hope that students will "catch up" to expectations. Existing, but limited, evidence finds students do make meaningful gains in achievement under this kind of remedial intervention. Yet the increase in quantity of math instruction is typically short-lived, suggesting an important but to-date largely unaddressed question: To what extent do students' gains from remediation persist after the quantity of instruction they receive returns to typical levels?

Some amount of decay in remediation-induced gains would not be a surprise. Several other short-lived educational interventions—assignment to a more skilled teacher, reducing class size, retaining students—show initially positive effects that begin to fade soon after the infusion of extra resources ends. Students' gains from extra class time may, nevertheless, fade more or less quickly than alternative approaches to boosting math skills. Teachers could, for example, use the extra time to broaden the curriculum, including concepts that better prepare students for future courses; or the extra time could be used for additional practice to reinforce the basic curriculum. Alternatively, an extra remedial class may simply be given over to test preparation boosting scores but not skills. Thus accurately judging the costs and returns to competing interventions, even if the produce similar initial gains, requires understanding the pattern of persistence over time.

In this paper I study a sample of middle-school students who were quasi-randomly assigned to either a math remediation class schedule—taking two math classes for one entire school year—or to a regular class schedule—taking one math class and one elective class in some other subject. I first estimate the immediate effect of taking an extra math class on student achievement at the end of the treatment year. I then estimate how much of that initial effect persists one and two years after students return to the regular one math class schedule. Finally, I examine treatment effects during high school years: math course taking and test scores, course taking in the subjects displaced by extra math time, and persistence to graduation.

Each year, starting with 6th grade, students in the Miami-Dade County Public Schools are identified as candidates for the two math class schedule if their score on the prior-spring's state math test falls below a pre-determined cut-score. Several factors determine students' final class schedules, but the probability of taking two math classes changes discontinuously and substantially at the cut-score. Using fuzzy regression discontinuity (FRD) methods, I estimate the local average treatment effect (LATE) comparing outcomes for treated students just below the cut to non-treated students just above. Since students were reassigned each school year, I explicitly model outcomes over-time using a dynamic treatment effects approach.

At the end of the school year during which they took two math classes, students who began the year with achievement near the 50th percentile (the assignment cut) scored 0.176σ (student standard deviations) higher than their otherwise identical classmates who attended just one math class. At a second discontinuity in the probability of treatment, about the 24th percentile of prior achievement, treated students gained 0.166σ in math. However, one year later, after a full year back on the traditional schedule of just one math class, the gains had shrunk to one-half to two-thirds the original size. Two years later the difference was one-fifth to one-third

the original gain. Once students reach high school I find little evidence of differences, though the estimates are comparatively less precise. For example, treated students were no more likely to have completed Algebra I by the end of 9th grade or to have completed Algebra II by the end of high school.

One key cost of a two math class approach is the opportunity cost of forgone instruction in some other subject (assuming the length of the school day is fixed). In Miami's middle schools, students taking two math classes gave up an elective class in physical education, the arts or music, or foreign languages. While the exact courses displaced in Miami are endogenous, the general constraint binds for any school taking this approach to remediation. First, this crowd out will directly reduce achievement in the forgone subjects during the short run (treatment year), which may alter the trajectory of longer run achievement. Second, assuming math is more cognitively taxing than the displaced course, the crowd out may also reduce effort in other classes, or reduce homework effort by increasing the marginal value of students' leisure time.

I find no treatment effect on reading test scores at the end of the treatment year, nor do I find strong evidence of any treatment-induced differences in outcomes during students' high school years. However, because fewer cohorts of students have aged through high school, those analysis samples are smaller and effect estimates much less precise. I cannot reject zero treatment effect for (nearly) all the high school outcomes examined, but I also cannot reject large benefits (losses) which would be important considerations in policymakers' calculus.

A handful of recent empirical papers, for example Chetty et al. (2011) and Chetty, Friedman, and Rockoff (2013b), document cases where initially large student test score gains decay substantially in the years immediately following treatment, but years later treatment effects re-emerge in longer-run, non-test-score outcomes. The decay of initial gains documented

in this paper closely mirrors the initial pattern of test score fadeout. However, if doubling math instruction in middle school has long-run effects, those re-emergent effects are not clearly detectable in students' high school achievement and attainment.

These estimates can be interpreted causally under three key assumptions. First, students' individual test scores, the forcing variable, and the treatment assignment cut-score were determined independently of each other. This assumption is well supported by the institutional details of the testing process, and by empirical tests. Second, no unobserved determinant of outcome test scores changed discontinuously at the assignment cut-score; the exclusion restriction in the instrumental variables terms of FRD. Other *observed* determinants did change discontinuously at the cut-scores. Students scoring just above the cut were slightly more likely to be placed in an honors or advanced section for their regular math class, and had higher-achieving, more-homogeneous classmates in their regular math class. I include these other "treatments" as additional endogenous variables, and use a vector of excluded instruments defined by interacting the standard FRD cut-score instrument with indicators for each individual school. This approach leverages between-school variation in how the cut-scores affect student assignment to math classes, both remedial and regular, for identification. The results from this multi-site method are similar to standard FRD estimates which ignore these other "treatments."

A third assumption is required when estimating the persistence of effects one and two years after treatment ends: namely that the immediate treatment effect of an extra math class does not depend on having been treated previously. I show that the results are similar under an alternative third assumption: that treatment in one year does not in practice affect the probability of treatment in future years.

Current school evaluation systems, which focus largely on student test scores in just a few subjects, have led many public schools to change the way they allocate scarce resources across subjects, like student time in each subject class (Dee and Jacob 2010). Unfortunately the literature on how quantity of instruction affects educational production is much less developed than the literatures for other inputs like teachers, peers, and class size. This paper contributes evidence to both that quantity literature generally, and the literature on remediation specifically.

1. Evidence on Quantity of Instruction and Persistence

1.1. Quantity of Instruction

Quantity of instructional time is an intuitive input to educational production, and has been the object of scholarship at least since psychologist John Carroll's work in the 1960's. The first order relationship is straightforward: the marginal returns to instructional time should be positive but diminishing. Most existing empirical evidence comes from settings where students experienced an increase in the total quantity of instructional time: longer school days, weeks, or years; summer school; or grade retention. The strength of causal inference varies but estimates are generally positive or null.[1]

This paper focuses on a much less studied type of variation in quantity of instructional time: holding total school hours fixed, but increasing instructional time in one subject by

[1] Checkoway et al. (2011) find, in general, no effects of a longer school day on achievement in reading, math, and science. Studying broader interventions which include longer school days, Hoxby, Murarka, and Kang (2009) find evidence of positive effects of longer days, but Angrist, Parthak, and Walters (2011) report no marginal effect. Studying variation in the number of school days, Sims (2008) finds positive effects on math achievement but not reading. Patall, Cooper, and Allen (2010) provide a review of older work on the length of school days and school years. Lavy (2010) finds positive effects using cross-country variation.

Jacob and Lefgren (2004) use regression discontinuity methods to estimate the effect of summer school and grade retention; they find positive effects on 3rd grade achievement, but not 6th grade. Schwerdt and West (2012) also find positive effects on students retained in 3rd grade using regression discontinuity. A broad review of the evidence finds positive effects for summer school (Cooper et al. 2000), but few of the studies reviewed have strong claims to causality.

reducing time in a different subject. The most comparable prior work comes from a policy in the Chicago Public Schools which doubled Algebra class time for low-achieving ninth graders. Using regression discontinuity methods, Raudenbush and Nomi (2013) find a gain of 0.31σ in test scores for treated students near the national median of math achievement (the point at which treatment was discontinuously assigned).[2] In a similar RD setting, Dougherty (2012) finds large gains in reading among 6th grade students assigned to take an extra reading class. These marginal gains from adding a second class in math (reading) are large, but somewhat smaller than descriptive estimates of a typical year of growth in math scores based on a national sample of students—students who are presumably taking just one math class (Hill et al. 2008). Additionally, Cortes, Goodman, and Nomi (forthcoming) find positive treatment effects on the Chicago students' high school graduation and college enrollment outcomes.

Other research has used between-country variation (Lavy 2010) or between-teacher variation (Brown and Saks 1987, Raudenbush, Hong, and Rowan 2002) in subject-to-subject time allocation decisions. The results are generally positive or null.[3]

1.2. Persistence of Achievement Gains

While proximate gains from increasing instruction may be an intuitive result, the extent to which those gains will persist over time is less clear ex-ante. Several studies over the past decade have documented the impersistence of student achievement gains generated by varying

[2] Original estimates from the Chicago policy were smaller, about 0.20σ (Nomi and Allensworth 2009, Raudenbush, Reardon, and Nomi 2012), but the more recent estimates account for differences in peer quality which was also discontinuously assigned at the same cut-score.

[3] Reallocating time across subjects is also often one element of broader bundle of school reforms (Bryk, Lee, and Holland 1993, Kemple, Herlihy, and Smith 2005, Lavy and Schlosser 2005). These broad treatment bundles often produce positive effects, but the marginal role of time allocation is unclear.

A related question is how to organize instructional time conditional on both total amount and subject allocations. Zepeda and Mayers (2006) review research on "block scheduling", a common alternative to traditional organization.

interventions. Indeed, across a number of quite different interventions, large initial gains decay about half or more in the first year after the intervention ends; similar to the estimates of fadeout I report in this paper. This result seems to hold for reducing class size (Krueger and Whitmore 2001), assigning students to more effective teachers (Kane and Staiger 2008, Jacob, Lefgren and Sims 2010, Rothstein 2010, Chetty, Friedman and Rockoff 2013b, Cascio and Staiger 2012), and retaining students in early grades (Jacob and Lefgren 2004, Schwerdt and West 2012). When researchers can measure persistence over multiple years, however, they generally find some persistence of the original gain.

(Im)persistence of achievement gains may arise from a number of mechanisms (Jacob, Lefgren and Sims 2010, and Cascio and Staiger 2012). First, students forget what they once knew. Students with different initial achievement nevertheless have similar rates of decay over time in math knowledge and skills (Bahrick and Hall 1991). Second, empirical estimates of learning persistence will exclude any real and persistent learning that is not measured on subsequent years' tests. A student may know lots of algebra one year after her algebra class but few algebra questions will appear on a geometry test. Third, schools' (families') may allocate educational resources endogenously over time in compensatory or complementary ways. A student assigned a low quality teacher one year may be ensured a high quality teacher the next. The similarity of estimated fadeout patterns across different interventions suggests the possibility that some or all of these causes will erode achievement gains no matter how they are obtained.[4]

Yet having more time with students allows for instructional strategies which may counteract mechanisms one and two. Extra classroom time could be used for extra individual

[4] Test scaling and cheating also influence empirical estimates of test score fadeout. Notably, Cascio and Staiger (2012) demonstrate that the typical practice of standardizing student test scores by cohort and year accounts for 10-20 percent of estimated fadeout. I return to this issue later. Jacob and Levitt (2003) document teacher cheating and show how it complicates estimation and interpretation of fadeout.

practice, as opposed to extra lecturing. Practice reduces the decay of knowledge and skills over time, and improves ability to apply existing knowledge to new concepts and problems (Glick and Holyoak 1983, Bahrick and Hall 1991, Ericsson, Krampe and Tesch-Romer 1993). Extra time might also be used to broaden the curriculum, including concepts that intentionally prepare students for future courses. Carrell and West (2010) show evidence that some teachers, perhaps consciously, improve their students' longer-run achievement at the expense of the short run achievement. Alternatively, any extra time may simply be given over to test preparation boosting scores but not skills. In the current data I do not observe how the extra time was used nor can I differentiate mechanisms of fadeout, but contrasting estimates of net persistence rates across interventions is a first step.

In general, empirical estimates of persistence for quantity of instruction induced gains are scarce. Jacob and Lefgren (2004) and Schwerdt and West (2012) both find initial positive effects on achievement of retention in 3rd grade, both find only about half of the initial effect persisting after one year treatment, and Schwerdt and West report additional decay in the years that follow.

To this literature I add estimates of both initial effect and effect persistence over time. I study a new setting, additional math instruction in middle school that crowds out other subjects, and apply estimation methods that explicitly model the dynamic nature of treatment effects.

2. Treatment and Setting

Each spring Florida students in grades 3-10 are tested in mathematics; scores are divided into five discrete ordered categories known as "achievement levels". By state statute, middle-school students (grades 6-8) whose scores the prior spring place them in level 1 or 2 are

identified as candidates for the remedial two math class schedule.[5,6] Several factors determine students' final class schedules; for example, a student's need for two reading classes may be greater than her need for two math classes, or her parents may intervene. Nevertheless, the probability of actually taking a two math class schedule does change discontinuously and substantially at the cut-scores separating achievement levels.

As shown in figure 1, in the Miami-Dade County Public Schools, students' assignment to a second, remedial math class was indeed strongly determined by their test score achievement level. Figure 1 plots the proportion of 6th grade students taking a second math class on the y-axis against math test scores the prior spring (their 5th grade year) on the x-axis. Consider the students who scored near the dashed line separating level 2 and level 3—about the 50th percentile of 5th graders in Miami. Only 2-3 percent of students who reached level 3 took two match classes. By contrast, their peers who barely missed level 3, but scored at the top of level 2, were about 13 percentage points more likely to take a second class. Notice that the probability of treatment jumps again, by about 19 points, at the cut between level 1 and level 2—about the 24th percentile in Miami. Figures for 7th and 8th grade assignment, available in the online appendix, show a similar pattern. Point estimates for the magnitude of the discontinuities at each cut-score in figure 1 are provided in row 1 of table 1.[7]

[5] § 1003.3156(1)(c) Florida Statute. This statute corresponds to the Florida Comprehensive Assessment Test (FCAT) which was the test during the period under study. The state is transitioning to the FCAT 2.0, but the requirement continues.

[6] In 2009 the score defining level 2 versus level 3 on the 8th grade math FCAT was roughly just above "Basic" on the NAEP; a result similar for 4th grade FCAT, and for both grades in 2005 (Banderia del Mello 2011). FCAT level 3 is "Proficient" in the No Child Left Behind sense.

[7] The estimates in table 1 are sharp regression discontinuity estimates. Column 1 of table 1 are the first stage estimates in the standard fuzzy regression discontinuity design. Each estimate is $\hat{\beta}$ from the specification

$$Y_{i,t} = \alpha + \beta \mathbf{1}\{A_{i,t-1} < c\} + \gamma A_{i,t-1} + \lambda\left(A_{i,t-1} * \mathbf{1}\{A_{i,t-1} < c\}\right) + \nu_{i,t},$$

where $Y_{i,t}$ is an indicator = 1 if the student took a class of the given type during 6th grade, $A_{i,t-1}$ is the forcing variable (5th grade math score), and c is the cut-score dividing achievement levels. I fit this specification with local-linear least squares using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score.

These discontinuities in the probability of taking a second math class are the core source of identifying variation. Figure 1 shows the average pattern of treatment assignment across all Miami-Dade middle schools. As I describe in more detail later in the paper, the magnitude of the treatment assignment discontinuities varies from school to school. This between school variation is also an important source of identification.

Students in the remedial schedule take two distinct math classes: a regular standard math class, which they would take in any case; and the second math class, usually called "Intensive Mathematics." An Intensive Math class enrolls, on average, 17.8 students (standard deviation 6.1) compared to a mean of 21.5 (6.3) in other regular math classes. In nearly all cases, a student's Intensive Math class is distinct from her regular math class. During the study period, about nine out of ten students in Intensive Math classes had a different teacher, or different peers, or both in their regular math class. Approximately half of Intensive Math classes were composed of only achievement level 1 students or only level 2 students, the other half mixed level 1 and level 2 students together.

The state's course description for Intensive Math early in the study period guided teachers to cover "mathematics content…identified by screening and individual diagnosis of each student's need for remedial instruction, critical thinking, problem solving, and test-taking skills and strategies." A later description is less specific, simply stating that the course is "designed to assist students with content mastery."[8] To the extent treated students' second math class met these individualized instruction objectives, the estimates I present later may be an upper bound for a more general policy of doubling the quantity of math instruction time.

Section 3 provides a more detailed description of the regression discontinuity methods used in this paper, including the selection of bandwith.

[8] Copies of the full course descriptions are available in the online appendix.

Devoting more of a student's school day to math instruction must crowd out other subjects, assuming total school hours remain fixed. Thus, the treatment condition is defined by both taking a second math class, and not taking a class in some other subject. In Miami, a second math class is most likely to displace a class in physical education, arts and music, or foreign language; though the specific class displaced is endogenous.

Figure 2, constructed just as figure 1 was, plots the proportion of students taking a class in English language arts (top line, triangles), physical education (diamonds), arts and music (circles), or foreign language (bottom line, squares). Plots for science and social studies classes, not shown, are quite similar to the plot for English language arts. Point estimates for the discontinuities in figure 2 are provided in table 1. For example, 60 percent of students who scored at level 3 took a P.E. class, but students who just missed the level 3 cut-score were 4 percentage points less likely to take a P.E. class. Discontinuities are also apparent in arts and foreign languages at the lower achievement level cut-scores.

These reductions in elective courses appear directly attributable to treatment assignment. Together the reductions in P.E., art, and foreign language course taking account for 75-80 percent of the displacement required by the increases in math classes. By contrast, there are no discontinuities in core subjects, English, science, and social studies; nor are there discontinuities in elective course taking at the higher achievement level cut-scores. Additionally, in results available in the online appendix, there are no discontinuities in 7th or 8th grade course taking at the 5th grade test cut-scores.[9,10]

[9] A difference in art and music course taking in 8th grade is statistically significant at the 10 percent level for students at the level 2/3 cut-score on the 5th grade test, but this is one significant result out of 24 tests (six subjects, two cut-scores, two grades).

[10] Attributing the discontinuities in P.E., arts, and foreign language course taking entirely to displacement by a second math class requires the assumption that the math achievement level cut-scores are not used in some other relevant class assignment mechanism. For example, the math score categories may be used in assigning students' science classes, which may in turn restrict which other courses would fit into a student's schedule. Nevertheless, two

This forgone instruction in P.E., arts, or foreign languages is a (potentially) important cost of the treatment in Miami, in addition to the direct costs of, for example, employing additional math teachers. In section 5 I examine the effects of treatment on outcomes in these displaced subjects. These crowd-out costs are, however, the endogenous choices of Miami's students and schools. In other jurisdictions, additional math time could displace different subjects, or lengthen the school day. Schools in Miami chose to reduce P.E., arts, and foreign language.

The data for this paper, provided by the Miami-Dade County Public Schools, are composed of administrative records on each student's annual state test scores, class and teacher assignments in each grade, and demographic information for the school years 2003-04 through 2012-13.

Miami-Dade is the fourth largest school district in the United States including, in any one year, about 80 thousand students in grades 6 through 8 at approximately 145 different middle schools. Table 2 column 1 provides some description of the district's middle school students during the study years. Nearly two-thirds are Hispanic and one-quarter African-American. Two-thirds live in households whose low income level qualifies them for free or reduced price lunch. A little over half have limited proficiency in English.

Column 2 of table 2 reports the same characteristics for my primary estimation sample: students who were (i) enrolled in a Miami-Dade school for all three middle school years, and (ii) have a math test score each of those three years. Comparing columns 1 and 2, this sample started 6th grade a few percentile points higher in the math distribution compared to their peers who

---

pieces of evidence support this assumption. First, the level 3/level 4 and level 4/level 5 cut-scores—which are not relevant to treatment assignment—show no discontinuities in elective course taking. Second, course enrollment forms provided to students explain that if they are assigned to a second math class, then they will not get to choose an elective course. Example forms are provided in the online appendix.

would eventually attrit, though both groups had similar growth in 5th grade. Non-attriting students may also have been less absent from school in 5th grade. The remaining columns of table 2 are discussed in the next section.

Focusing on these non-attriting students keeps the sample constant as I estimate effects over time, though, as I show later, results are robust to using more inclusive samples. Additionally, in results available in the online appendix table A1, I find no evidence of differential attrition rates using traditional tests. Students who score just below the level 1/level 2 cut-score or just below the level 2/level 3 cut-score are no more or less likely to leave the Miami-Dade schools, have a missing test score, or be retained in grade than those scoring just above the cut-scores.

3. Immediate Effects on Math Achievement

3.1. Estimation Methods

The first empirical objective is to estimate the effect of taking a second remedial math class, in addition to one's regular math class, on student achievement at the end of the treatment school year. As figure 1 shows, the probability of treatment changes discontinuously at the level cut-scores. Thus, for students scoring sufficiently close to the cut-score, the remediation treatment is effectively randomly assigned by test measurement error, but compliance with treatment is imperfect. Under certain assumptions, fuzzy regression discontinuity (FRD) methods recover the local average treatment effect (LATE) in such settings; where "local" is limited to compliant students scoring at (or at best very near) the test cut-score.

The estimand of interest, δ_t, is defined in terms of limits:

$$\delta_t \equiv \frac{\lim_{A_{i,t-1}\uparrow c} E\left[A_{i,t}|A_{i,t-1}\right] - \lim_{A_{i,t-1}\downarrow c} E\left[A_{i,t}|A_{i,t-1}\right]}{\lim_{A_{i,t-1}\uparrow c} E\left[R_{i,t}|A_{i,t-1}\right] - \lim_{A_{i,t-1}\downarrow c} E\left[R_{i,t}|A_{i,t-1}\right]}$$

where $A_{i,t}$ is the math state test score for student i in year t, $R_{i,t}$ is the treatment indicator $= 1$ if the student had a second, remedial math class, and c is the proficiency level cut-score.

I estimate δ_t using the local linear instrumental variables regression method suggested by Hahn, Todd, and Van der Klaauw (2001) and Imbens and Lemieux (2008). Specifically, I fit the regression equations:

$$A_{i,t} = \alpha + \delta_t R_{i,t} + f\left(A_{i,t-1}\right) + g\left(A_{i,t-1}\right) * \mathbf{1}\left\{A_{i,t-1} < c\right\} + \varepsilon_{i,t} \tag{1}$$

by two-stage least squares instrumenting for the treatment indicator $R_{i,t}$ with an indicator for scoring below the cut-score $\mathbf{1}\left\{A_{i,t-1} < c\right\}$. I estimate equation 1 and using only observations where $\left|A_{i,t-1}\right| \leq (c + b^*)$, where b^* is the optimal smoothing parameter, the bandwidth, determined using the cross-validation procedure described in detail by Ludwig and Miller (2007) and Imbens and Lemieux (2008).[11] While b^* is optimal given a mean squared prediction error criterion, I show that the main results are robust to bandwidth choice.

Both f and g are functions only of the forcing variable $A_{i,t-1}$, the math test score whose achievement level determines treatment. Given the strong empirical linear relationship between

[11] Throughout the paper I report estimates with $b^* = 23$ which is the optimal bandwidth for the key first stage equation (i.e., predicting compliance with treatment in 6th grade).

In brief, the cross-validation procedure chooses the bandwidth that minimizes the mean squared prediction error

$$b^* = \arg\min_b \frac{1}{N}\sum_{i=1}^{N}\left(Y_{i,t} - \hat{Y}\left(b, A_{i,t-1}\right)\right)^2$$

where $\hat{Y}\left(b, A_{i,t-1}\right)$ is the out-of-sample predicted value for some outcome $Y_{i,t}$ using parameter estimates from a simple bivariate regression of $Y_{j,t}$ on the forcing variable $A_{j,t-1}$ and a constant using only observations $j \neq i$ where $\left(A_{i,t-1} - b\right) \leq A_{j,t-1} < A_{i,t-1}$ when $A_{i,t-1} < c$, and where $A_{i,t-1} \leq A_{j,t-1} < \left(A_{i,t-1} + b\right)$ when $A_{i,t-1} \geq c$. In this case, I choose b^* using the first-stage equation, $Y_{i,t} = R_{i,t}$, which is (nearly) always less than the optimal bandwidth suggested by the ITT equation, $Y_{i,t} = A_{i,t}$.

student test scores over time, the results presented in this paper use linear functions for f and g, allowing the two slopes to differ. The results are robust to using higher order polynomials. In all estimates I adjust the standard errors to account for the coarse nature of the forcing variable, $A_{i,t-1}$, by clustering observations at each scale score point (Lee and Card 2008).

Causal identification of δ_t requires two important assumptions. First, that treatment assignment is independent of potential outcomes, $A_{i,t}$, conditional on the forcing test score, $A_{i,t-1}$. More generally, the cut-score, c, and the forcing test scores, $A_{i,t-1}$, are determined independently of each other. This seems a reasonable assumption. Cut-scores and proficiency levels are determined in "scale score" units, not the raw number of correct items; scale scores are a function of IRT (item response theory) estimated parameters unknown to students taking the exams. Additionally, the data do not show any evidence against independence. Figure 3, a histogram for 6th grade students of the forcing test score, $A_{i,t-1}$, shows a smooth distribution near the cut-scores. Using the formal test suggested by McCrary (2008) I fail to reject the null hypothesis of continuity at all four cut-scores; the discontinuity estimates and standard errors are -0.031 (0.023) at the level 2 cut-score, 0.010 (0.015) at level 3, -0.017 (0.015) at level 4, and -0.015 (0.030) at level 5.

Second, except the probability of treatment, nothing that affects the outcome, $A_{i,t}$, changes discontinuously at the cut-score. This assumption amounts to the exclusion restriction in fuzzy regression discontinuity setting. There is no evidence this assumption is violated for any pre-treatment observables. Figure 4, following the structure of figure 1, plots math test score gains in year $(t-1)$, the proportion eligible for free or reduced price lunch, and school absences in the year $(t-1)$ against the forcing test score, $A_{i,t-1}$. Columns 5 and 6 of table 2 report point

estimates of the potential discontinuity at the cut-scores for several pre-treatment variables; none of the estimates is statistically significant.[12]

There is, however, evidence that other educational inputs in treatment year t were discontinuous at the achievement level cut-scores. As shown in figure 5, the characteristics of students' regular required math class were somewhat different for students scoring just above and just below the cut-scores. Students scoring just above an achievement level cut-score were more likely to be in an honors or advanced section for their regular math class, and their regular class peers' prior math achievement was higher (mean) and more homogeneous (standard deviation). By contrast, I find no evidence of a discontinuity in the characteristics of students' regular math teacher, including teacher test-score value-added, years of experience, or having a master's degree.[13] Point estimates for the discontinuities in figure 5, and for (the lack of) discontinuities in teacher characteristics are provided in table 3.[14]

To the extent these other discontinuously-assigned "treatments" affected math achievement my estimates will be biased, but the direction of bias is unclear. Existing theory and empirical evidence suggests that students learn more when placed with higher achieving, more homogenous peers (Lazear 2001, Hoxby 2002, Duflo, Dupas, and Kremer 2011) which would bias against finding a treatment effect. By contrast, achievement can suffer when students are assigned to courses beyond what they are prepared for (Clotfelter, Ladd, and Vigdor 2011) suggesting a potential positive bias.

[12] Table 2 columns 5 and 6 are estimated with standard sharp regression discontinuity methods. See footnote 7 for the exact specification.

[13] Teacher test-score value-added measures were estimated following the method described in Kane and Staiger (2008), but are not corrected for measurement error (i.e., "shrunken") since they are an outcome variable in the present analysis. Additionally, for student i assigned to teacher j in year t, is the measure is teacher j's value-added estimated using data from all years 2004-05 through 2012-13 but excluding year t.

[14] Table 3 is estimated with standard sharp regression discontinuity methods. See footnote 7 for the exact specification.

To address the potential bias I add these three regular math class characteristics as additional endogenous treatment variables to the specification described by equation 1, which becomes:

$$A_{i,t} = \alpha + \delta_t R_{i,t} + C_{i,t}\beta + f\left(A_{i,t-1}\right) + g\left(A_{i,t-1}\right) * \mathbf{1}\{A_{i,t-1} < c\} + \phi_s + u_{i,t} \quad (2)$$

where the vector $C_{i,t}$ includes an indicator for assignment to an honors or advanced section for regular math class, plus both the mean and standard deviation of baseline math achievement, A_{t-1}, among classmates in student i's regular math class. As excluded instruments I use a vector composed of interactions between an indicator for each school and the below-cut-score indicator, $\mathbf{1}\{A_{i,t-1} < c\}$. I estimate equation 2 using limited information maximum likelihood; LIML is preferable to two-stage least squares in a setting, like this, with many instruments derived from the multi-site structure of treatment assignment (Chamberlain and Imbens 2004). This multi-site multi-treatment IV method has been used in other settings by Kling, Liebman, and Katz (2007) and Duncan, Morris, and Rodrigues (2011). Results from the more-typical one-instrument one-treatment FRD approach are quite similar, and are presented alongside this preferred approach in table 5.

This approach uses between-school variation in how the cut-scores affect student assignment to math classes, both remedial and regular, for identification. When examined school-by-school the size of the discontinuities in $R_{i,t}$ and $C_{i,t}$ do vary, and between-school variation has been demonstrated in other class assignment decisions (Clotfelter, Ladd, and Vigdor 2005, 2006, Loeb, Kalogrides, and Beteille forthcoming). Table 4 describes the between-school variation in how the two cut-scores are used in treatment assignment. For each of the 145 schools, I estimated the school-specific discontinuity in each treatment—two math classes,

advanced or honors section, class baseline score mean, class baseline score standard deviation—at each 5th grade test cut-score. Estimation was identical to the methods used in tables 1 and 3, except that the sample was restricted to only student observations from the given school.

Table 4 columns 1 and 2 show the meaningful between-school variation in how the two math class treatment are assigned. In the top decile of schools the take-up rates are three times as large as the district average, and in many schools there is effectively no discontinuity. The latter result is partly because in one out of four schools no level 2 students are assigned to an extra math class. Additionally, schools with large assignment discontinuities at the level 1/2 cut tend to have smaller discontinuities at the level 2/3 cut; the two estimates are correlated -0.39. There is also meaningful between-school variation in how students' regular math class characteristics are assigned at the cut-scores, but the discontinuities in these other treatments are only weakly correlated with the discontinuities in the second math class treatment assignment.

While the multi-site FRD estimator requires, and indeed benefits from, differential treatment take-up between schools, I assume that treatment effects are constant. If treatment effects also vary between sites, multi-site instrumental variables estimates will be biased if the site-by-site treatment compliance rates covary with the site-by-site treatment effects (Reardon and Raudenbush 2013). Reardon and coauthors describe this "compliance-effect covariance bias" and related issues, and develop estimators for such settings (Raudenbush, Reardon, and Nomi 2012, Reardon and Raudenbush 2013, Reardon et al. 2014).

Additionally, as Reardon and Raudenbush (2013) show, the multi-site, multi-treatment method requires an additional assumption beyond the standard IV assumptions to identify δ_t: namely that assignments $C_{i,t}$ cannot be affected by $R_{i,t}$ conditional on the excluded instrument, $\mathbf{1}\{A_{i,t-1} < c\}$. While figure 5 and table 3 show evidence the instrument affected $C_{i,t}$, this

assumption would be violated only if assignment to take a second math class had an additional effect on $C_{i,t}$. This assumption seems plausible. First, students' remedial and regular math classes were separate assignments, as described in section 2. Students with one math class and students two math classes were mixed together in regular math classes, and students from different remedial classes were mixed in regular math classes. This mixing suggests a student's regular math class was not directly determined by her assignment to a remedial math class. Second, there are similarly sized discontinuities in $C_{i,t}$ at the level 3/level 4 cut-score where $R_{i,t}$ is not relevant.

3.2. Effects at the End of the Treatment Year

The top panel of figure 6 plots the average test score change from 5th to 6th grade, measured in student standard deviation units, by the scale score values of the forcing variable, students' 5th grade math score $A_{i,t-1}$.[15] At the end of the treatment year, 6th grade, there is a visible discontinuity in student test score gains at both the level 1/level 2 and level 2/level 3 cut-scores, consistent with a positive effect of additional math class time. Students scoring just below the level 1/level 2 cut gained, on average, about 0.10σ compared to 0.05σ for students just above the cut. The apparent gap at the level 2/level 3 cut is similar.

However, the differences depicted in figure 6 understate the effect of taking a second, remedial math class since only the probability of treatment was discontinuous at the cut-scores (figure 1). Table 5 reports estimates of the effect for students who were actually treated—specifically, the effect for students who took a second class because they scored just below the cut (the LATE). Students who began 6th grade near the 50th percentile of math achievement (the

[15] With the notable exception of forcing variable, $A_{i,t-1}$, all test scores in this paper have been standardized (mean 0, sd 1) using the grade-by-test-year distribution. Test score gains in figure 6 are relative to the pre-treatment test score, i.e., $(A_{i,t} - A_{i,t-1})$ for the top panel, $(A_{i,t+1} - A_{i,t-1})$ for the middle panel, and $(A_{i,t+2} - A_{i,t-1})$ for the bottom panel.

level 3 cut) and who took a second math class during the year scored 0.176σ higher at the end of the year (panel A, row 1, col 5). The gain among treated students who began 6th grade near the 24th percentile was quite similar: 0.166σ (panel A, row 1, col 1).[16] In short, these estimates suggest that taking two math classes, one remedial and one traditional, raises median students' math achievement by one-sixth or more of a standard deviation beyond the gain from taking just one, traditional class.

These treatment year gains are substantively important. The 0.16-0.18σ effect is similar in magnitude to the oft-estimated standard deviation in teacher effectiveness (Hanushek and Rivkin 2010), and the effect of smaller class-size in early grades (Krueger 1999). Recent work suggests differences in test scores during elementary and middle grades can predict long-run labor market outcomes (Chetty, Friedman, and Rockoff 2013b). Yet, while the variation in class time was binary (one class to two), the results are consistent with diminishing marginal returns to class time. A 0.16-0.18σ gain is only one-third to two-thirds the size of what Hill et al. (2008) estimate as the typical growth in math test scores during middle-school; Hill et al. use data from national norming samples of students, most of whom presumably took just one math class.

Notably, the estimates in table 5 measure achievement gains in math skills that are tested at the end of 6th grade. The two-math-class treatment has a remedial education motivation, and thus may have had larger effects on the kinds of math skills students should have learned in earlier grades. If students were re-tested on 5th grade level material, or earlier material, the estimated treatment effects on those math skills might even be larger.

To this point and throughout the rest of the paper I focus discussion of results on effects for students treated in 6th grade who I can subsequently observe in both 7th and 8th grade, as in

[16] As suggested by figure 1, the excluded instrument(s) are strong predictors of treatment. The *F*-statistics reported in table 5 are all well above the standard threshold.

figure 6. This choice improves the estimation and interpretation of effects over time, the subject of the next section. However, estimates of the immediate effects are similar for students treated in 7th or 8th grade as shown in panels B and C of table 5. The estimates are also robust to attrition from the original 6th grade cohorts. As a comparison to row 1, in row 2 I add to the estimation sample students who attrit after 7th grade. In row 3 I add students who attrit after 6th grade. Neither changes the estimated effect substantially.

Three other tests of robustness are important to note. First, my preferred estimates were obtained using the multi-site FRD approach described above for fitting equation 2. This approach controls for the observable characteristics of students' regular math class that changed discontinuously at the assignment cut-scores (figure 5). Failure to account for these other educationally productive inputs risks biasing the estimates of the effect of a second math class; though, as discussed above, the direction of bias is ambiguous. For comparison columns 2 and 6 of table 5 report estimates obtained by traditional FRD methods ignoring the other "treatments". These estimates suggest a potential positive bias from omitting the regular math class characteristics. I cannot, however, reject the null of equality across the two methods, partly because the traditional FRD estimates are less precise.

Second, as an alternative test for bias from omitting the regular math class characteristic "treatments", I conduct a placebo test based on the achievement level 3/level 4 cut-score. Students near the level 4 cut-score were not subject to the second math class requirement, as shown in figure 1, but the regular math class characteristics were somewhat discontinuous at the level 4 cut as shown in table 3. Any differences in math achievement outcomes at level 4 provide one measure of the potential effect of these other "treatments." The results, not presented here,

suggest a small positive or zero effect on year t math score for students scoring just below the level 3/4 cut-score suggesting potential negative bias in the estimates of interest for this paper.

Third, my preferred estimates were obtained using the cross-validation optimal bandwidth which minimizes mean-squared prediction error, but the results are fairly robust to a wide range of bandwidth choices as shown in figure 7. Robustness to bandwidth choice is a critical test for the local linear regression methods used in this paper (Imbens and Lemieux 2008). Figure 7 plots a series of point estimates one for each integer bandwidth. The solid line traces the point estimates, and the dotted lines trace the cluster-adjusted 95 percent confidence intervals. As figure 7 shows, the estimates of immediate effects are generally not sensitive to bandwidth choice. For the level 2/level 3 cut-score, the estimates do get somewhat smaller at smaller bandwidths but they are also less precisely estimated.[17]

These LATE estimates are only appropriate for causal inference about a distinct population: middle-school students (i) with math achievement levels similar to the 24th or 50th percentile of the Miami-Dade distribution (i.e., near the cut-score); and (ii) who, under a Florida-like policy, would take a second math class (or not) based solely on their prior test score (i.e., compliers). While we cannot observe this population directly, table 2 provides some descriptive insight into the compliers. Columns 3 and 4 report the ratio of estimated compliance, from the single-instrument FRD first-stage, among students who share the given characteristic (e.g.,

[17] One additional test of bias is based on the between-school variation in how these other regular math class "treatments" were assigned. I estimate the treatment effect of interest using conventional single-instrument FRD methods but with a restricted subsample of schools—schools where characteristics of students' regular math class do not change discontinuously at the cut-scores. The estimates for this subsample are somewhat larger, suggesting a negative bias, but I cannot reject that they are equivalent to the main results. To identify the subsample I estimate three parameters for each school, the data underlying table 4. Then I select into the subsample any school where I cannot reject the null hypothesis of no discontinuity in each of the three characteristics. A stronger alternative, and perhaps preferable, definition of the subsample would include only schools where the three confidence intervals all include zero but also exclude the estimated district average. The subsample of such schools is too small to support reasonable estimation.

among female students only) over estimated compliance among all students. This ratio can be interpreted as the relative likelihood a complier has the given characteristic (Angrist and Pischke 2009, pp. 171-172). The compliers to whom we can make inference are, this evidence suggests, more likely to be female and limited in English proficiency. For students near the 50th percentile, compliers are also more likely to be African-American. Among students near the 24th percentile, compliers are more likely to be Hispanic and frequently absent the prior year.

4. Persistence of Math Gains Over Time

While adding a second math class appears to have substantial immediate effects on achievement, those improvements relative to one's peers may not, for reasons detailed earlier, persist over time. In this section my objective is to estimate the persistence of gains induced by a second, remedial math class one and two years after the extra class has ended.

Returning to figure 6, the middle panel plots the average test score gain from 5th to 7th grade, measured in student standard deviation units, by the scale score values of the forcing variable, students' 5th grade math score $A_{i,t-1}$. This visual evidence suggests achievement gains seen at the end of 6th grade may have substantially faded by the end of 7th grade. The bottom panel of figure 6, showing test score gains from 5th to 8th grade, suggests even further fadeout by the end of 8th grade.

In table 6 panel A I report LATE estimates for treatment in 6th grade on test score outcomes at the end of 7th and 8th grade. Column 1 repeats the 6th grade test score effect from table 5 column 1 row 1 for convenience. The 7th and 8th grade LATE are estimated using equation 2 and the multi-site multi-treatment FRD methods described earlier; the only difference between columns 1-3 in panel A is the choice of outcome variable.

Students who began 6th grade near the 50th percentile of math achievement (the level 2/level 3 cut-score) and who took a second math class during 6th grade scored 0.099σ higher at the end of 7th grade. The ratio of the 7th grade estimate and 6th grade estimate implies 56 percent of the initial gain persisted after one year. Put differently, 43 percent of the initial gain faded out. The persistence rate is somewhat higher, 84 percent, for treated students who began 6th grade near the 24th percentile (level 1/level 2 cut-score). By the end of 8th grade only 30-40 percent of initial gains persist on average.

However, the estimates in table 6 panel A (and the middle and bottom panels of figure 6) may either under- or over-state the true *marginal* effect of treatment in 6th grade. The ambiguity arises because students' treatment status—taking two math classes versus taking just one math class—is reassigned in 7th grade and reassigned again in 8th grade; and, since treatment assignment is based on prior math achievement scores, assignment to remediation in 7th (8th) grade is partly a function of remediation treatment in 6th grade. The estimates in table 6 panel A take no account of this dynamic treatment assignment, and thus represent the *total* effect or reduced form effect of treatment in 6th grade on achievement in 7th (8th) grade. If, for example, taking two math classes in 6th grade reduces the probability of taking two math classes in 7th (8th) grade, these reduced form estimates will understate the marginal effect of treatment in 6th grade on 7th (8th) grade achievement.

4.1. Estimation with Dynamic Treatment Assignment

To recover the *marginal* effect of treatment I apply an estimator which explicitly accounts for the dynamic nature of treatment assignment. The parameter of interest is the marginal effect of taking two math classes in year t on student achievement in a subsequent year

τ; specifically, the effect of treatment in year t but no treatment in any subsequent year $(t+1) \ldots \tau$.

$$\delta_{t,\tau}^{M} \equiv \frac{\partial A_{i,\tau}}{\partial R_{i,t}}, \qquad \tau \in \{t+1, t+2, \ldots\}$$

The implied persistence of remediation-induced gains from t to τ is given by the ratio $\delta_{t,\tau}^{M}/\delta_t$.

Because of the dynamic nature of treatment—assignment to remediation in any one year is partly a function of assignment to remediation in prior years—$\delta_{t,\tau}^{M}$ cannot be directly estimated by the typical FRD method. To see why consider equation 3 which decomposes the *total* effect of remediation in year t on achievement in year τ. For simplicity, equation 3 specializes to the case where $\tau = (t+1)$.

$$\delta_{t,t+1}^{T} \equiv \frac{\mathrm{d}A_{i,t+1}}{\mathrm{d}R_{i,t}} = \frac{\partial A_{i,t+1}}{\partial R_{i,t}} + \left[\frac{\mathrm{d}R_{i,t+1}}{\mathrm{d}R_{i,t}} * \frac{\partial A_{i,t+1}}{\partial R_{i,t+1}}\right] = \delta_{t,t+1}^{M} + \left[\pi_{t,t+1} * \delta_{t+1,t+1}^{T}\right] \tag{3}$$

As the decomposition makes clear, remediation treatment in year t can affect achievement in year $(t+1)$ directly, $\delta_{t,t+1}^{M}$, or by impacting future treatment assignment, $\pi_{t,t+1}$, or both. A naïve FRD approach regressing $A_{i,t+1}$ on $R_{i,t}$ in the manner of equation 2 would yield $\hat{\delta}_{t,t+1}^{T}$ not $\hat{\delta}_{t,t+1}^{M}$. Note that δ_t, which was the focus of section 3, is simply $\delta_{t,t}^{T}$ in this notation.

Equation 3 suggests a form of dynamic treatment effects estimator for $\delta_{t,t+1}^{M}$, first proposed by Cellini, Ferreira, and Rothstein (2010) for the regression discontinuity setting.[18] By rearranging equation 3 we have:

$$\delta_{t,t+1}^{M} = \delta_{t,t+1}^{T} - \pi_{t,t+1}\delta_{t+1,t+1}^{T} \tag{4}$$

[18] This is the "recursive" estimator in Cellini, Ferreira, and Rothstein (2010). The authors also propose an alternative "one-step" estimator to address imprecision arising because they are interested in effects at τ much more distant than $(t+1)$ or $(t+2)$.

The three right hand side terms can be estimated individually by the FRD methods described for equation 2, with $\delta^{M}_{t,t+1}$ then estimated in the style of indirect least squares. The estimates presented in table 6 panel B and discussed below follow this "indirect FRD" approach where year t is students' 6th grade year and year $(t+1)$ is 7th grade. Standard errors obtained by the delta method.[19]

Causal identification of $\delta^{M}_{t,t+1}$ requires three assumptions: the two standard FRD assumptions discussed in section 3, and a further third assumption that $\delta^{T}_{t+1,t+1}$ does not depend on prior treatment status, $R_{i,t}$.[20] Evidence for and against assumptions one and two is detailed in section 3. While that discussion focuses on the 6th grade assignment discontinuity, the evidence for the 7th and 8th grade discontinuities is similar and presented in the appendix. Moreover, to address the potential violations of assumption 2 seen in students' regular math class characteristics, I use the multi-site FRD approach to estimate each of the terms in 4.

The new third assumption would be violated if a second, remedial math class in 7th grade was less (more) effective for students who also had a second class in 6th grade. In other words, if there are diminishing (increasing) returns to each year with a second class. Note, however, that I do not constrain effects to be equal across grade levels; that is, δ_t is not necessarily $= \delta_{t+1}$. As one test of this assumption I first estimate $\delta^{T}_{t+1,t+1}$ using only the subsample of students where $R_{i,t} = 1$, second similarly estimate for $R_{i,t} = 0$, and finally test the null that the two estimates

[19] To obtain the complete variance/covariance matrix I simultaneously estimate the system of four equations, one each for the three right hand side terms in 4 and one for $\delta^{T}_{t,t}$. Each of the four equations is of the form equation 2. I stack observations, interact all right hand side terms with an indicator for the equation, include equation fixed effects, and estimate by the local-linear LIML methods described in section 3. This provides point estimates identical to LIML equation by equation, but with the added cross-equation covariances of coefficient estimates. This parallels the approach used by Cellini, Ferreira, and Rothstein (2010).

[20] The third assumption is not required for estimating $\delta^{T}_{t,t}$ in section 3 with the estimation sample is restricted to 6th grade students. No students are exposed to the remediation treatment in question before 6th grade.

are equivalent.[21] I cannot reject this null hypothesis. For students near the level 1/level 2 cut-score the test p-value is 0.250; near the level 2/level 3 cut-score the p-value is 0.525. However, this is an imperfect test since each of the subsample estimates is its own particular LATE.

To this point I have focused on describing estimation of $\delta^M_{t,t+1}$: the marginal effect of remediation in 6th grade on student achievement measured at the end of 7th grade. Extending the same logic, I also estimate $\delta^M_{t,t+2}$: the marginal effect of remediation in 6th grade on achievement at the end of 8th grade.[22] Equation 4 becomes:

$$\delta^M_{t,t+2} = \delta^T_{t,t+2} - \pi_{t,t+1}\delta^M_{t+1,t+2} - \pi_{t,t+2}\delta^T_{t+2,t+2} \tag{5}$$

I use the same methods and maintained assumptions to estimate $\delta^M_{t+1,t+2}$ as for $\delta^M_{t,t+1}$. Here the more plausible third identifying assumption requires that neither $\delta^T_{t+2,t+2}$ nor $\delta^M_{t+1,t+2}$ depend on $R_{i,t}$. Extending the partial test described earlier, I cannot reject the joint null of both equivalence of treatment effects in 7th grade and equivalence in 8th grade.[23] The p-values are 0.501 for the level 1/level 2 cut-score, and 0.326 for the level 2/level 3 cut-score.

4.2. Persistence During Middle School

Table 6 panel B reports estimates of the marginal effects, and implied measures of persistence, estimated using the indirect FRD methods described above for equations 4 and 5 to account for the dynamic nature of treatment assignment over time. Recall that students who

[21] I obtain these two estimates and their covariance by fitting equation 2 with $A_{i,t+1}$as the outcome, $R_{i,t+1}$ as the treatment, and $\mathbf{1}\{A_{i,t} < cut\}$ as the instrument; and with all right hand side variables interacted with the indicator $R_{i,t}$.

[22] I use the same "stacked" equation method described for $\delta^M_{t,t+1}$. In this case there are eight equations. Each of $\delta^T_{t,t}$, $\delta^T_{t,t+2}$, $\pi_{t,t+1}$, $\pi_{t,t+2}$ and $\delta^T_{t+2,t+2}$ require one equation; $\delta^M_{t+1,t+2} = \delta^T_{t+1,t+2} - \pi_{t+1,t+2}\delta^T_{t+2,t+2}$ requires three equations just as $\delta^M_{t,t+1}$.

[23] Specifically I test the joint null hypothesis:
H_0: $\left(\delta^T_{t+2,t+2}|R_{i,t}=1\right) = \left(\delta^T_{t+2,t+2}|R_{i,t}=0\right)$ & $\left(\delta^M_{t+1,t+2}|R_{i,t}=1\right) = \left(\delta^M_{t+1,t+2}|R_{i,t}=0\right)$.

began 6th near the 50th percentile of math achievement and who took a second math class during 6th grade had an immediate gain of 0.176σ. By the end of 7th grade that initial gain had shrunk by a little more than half to 0.077σ. By the end of 8th grade the difference was just 0.031σ, suggesting three-fifths of the initial achievement gains had been lost.

For students who began 6th grade near the 24th percentile the initial effects of treatment, which were quite similar to those at the 50th percentile, decayed somewhat less after one and two years. However, these estimates are less precise. I cannot reject complete persistence one year after treatment, and, conversely, I cannot reject complete fadeout two years after treatment. In short, the initial gains from a second, remedial math class, while substantial, do not fully persist if students return to a regular schedule with just one math class.

This pattern of fadeout over time is quite similar to the pattern reported in the literature for other interventions to improve achievement scores, like reducing class size or improving the effectiveness of teachers, described in section 1. For example, several studies now show that half, or less, of teacher-induced gains persist after one year (Kane and Staiger 2008, Jacob, Lefgren, and Sims 2010, Rothstein 2010). Chetty, Friedman, and Rockoff (2011) and Cascio and Staiger (2012) find similar one-year decay, but also track persistence for a number of years; they estimate that one-quarter to one-third of teacher-induced gains persist in the long run. This is similar to the pattern reported in table 6. The extent and regularity of impersitence patterns for this intervention—two math classes—and other interventions reinforce the importance of considering fadeout in school policy and management decisions.

In the end the marginal effect estimates (table 6 panel B) are fairly similar to the total effect or reduced form estimates (panel A), especially for students near the level 1/level 2 cut-score. Why the similarity? One possibility is that treatment assignment in year $(t + 1)$ does not

in practice depend (substantially) on treatment assignment in year t. If $\pi_{t,t+1} = 0$ in equation 4 then the total effect is equal to the marginal effect, $\delta^{M}_{t,t+1} = \delta^{T}_{t,t+1}$.[24] On its face $\pi_{t,t+1} = 0$ seems implausible. Section 3 and figure 6 provide evidence that remediation created a discontinuity in students' test scores at the end of 6th grade, $A_{i,t}$. Since $A_{i,t}$ partly determines $R_{i,t+1}$, we would intuitively expect a discontinuity in $R_{i,t+1}$. Yet the magnitude of $\pi_{t,t+1}$ is less easy to predict since it depends on several factors, including treatment assignment compliance.

Importantly, the magnitude of $\pi_{t,t+1}$ depends not simply on math scale scores, $A_{i,t}$, rather it depends on the "achievement level" category into which $A_{i,t}$ falls. Even if treatment in 6th grade improves student test scores from 5th to 6th grade, treatment may nevertheless have little or no effect on students' 6th grade "achievement level" if the 6th grade cut-scores are much higher (lower) in the distribution than the 5th grade cut-scores. For example, consider students whose 5th grade test score placed them just below the 24th percentile of math achievement; their 5th grade achievement level was level 1. By the end of 6th grade taking two math classes had boosted treated students' scores to the 29th percentile, but the minimum cut-score for achievement level 2 was at the 39th percentile, much higher than it had been for the 5th grade test. Figure 8 plots the probability of scoring in achievement level 1 on the 6th grade math test (top panel) and similarly the probability of scoring in level 2 (bottom panel) against 5th grade math test score. There is no evidence that treatment in 6th grade affected achievement level on the 6th grade test. Moreover, as shown in figure 1, not all level 1 and level 2 students were treated, so effect of treatment in 6th grade on the probability of treatment in 7th grade, $\pi_{t,t+1}$, is a fraction of any discontinuity that there might appear to be in figure 8. The pattern depicted in

[24] Under the assumption $\pi_{t,t+1} = 0$ the persistence estimates in table 6 panel A are closely related to the general persistence estimator proposed by Jacob, Lefgren, and Sims (2010). The assumption $\pi_{t,t+1} = 0$ is analogous to the exclusion restriction in Jacob, Lefgren, and Sims' instrumental variables approach.

figure 8 is similar when the y-axis is switched to achievement level on the 7th grade test (available in the online appendix).

Table 7 and figure 9 provide additional robustness checks for the persistence estimates. First, the main estimates, replicated in table 7 row 1, use a sample of non-attriting students. Excluding attriters would positively bias those estimates if, as seems plausible, the achievement growth of attriting students was less well served by treatment. To check for attrition related bias, row 2 of table 7 replicates the preferred estimation strategy but with a larger sample that includes students who attrit after 7th grade. With these attriters included, the one-year persistence point estimate is actually higher, contrary to the expected bias, though I cannot reject the null that the row 1 and row 2 estimates are equivalent. Second, in row 3 I report persistence estimates using the more common single-instrument FRD methods. These estimates are much lower, suggesting more bias here than in the estimates of immediate effects in table 5. However, again, I cannot reject the null that row 3 and row 1 are equivalent. Third, figure 9 shows the preferred estimation approach for varying bandwidths. Estimated persistence is robust to bandwidth choice, though less precisely estimated at smaller bandwidths. At the level 2/level 3 cut-score persistence two years out is more sensitive but also much noisier at small bandwidths.

Finally, Cascio and Staiger (2012) provide both theoretical and empirical evidence that estimated fadeout is partly a statistical artifact of the common practice, which I follow in this paper, of standardizing student test scores within grade and cohort. They estimate that perhaps 10-20 percent of estimated fadeout is simply an artifact of rescaling test scores each year to have unit variance while the true variance is growing. While this would negatively bias my persistence estimates, the potential bias is small relative to the effects reported here and I would still reject the null of complete persistence.

4.3. Math Outcomes During High School

While the initial math test score gains in 6th grade largely fadeout by the end of 8th grade, this fadeout does not rule out treatment benefits in the longer run. Chetty et al. (2011) and Chetty, Friedman, and Rockoff (2013b), as examples, document cases where large initial test score gains for young students fadeout, but in the long run treatment boosts college going, earnings, and other adult outcomes. I do not have measures of adult outcomes, but I do observe various attainment and achievement outcomes during high school. In the remainder of this section I report on high school math outcomes, and in the next section turn to outcomes in other subjects.

Table 8 columns 1 and 5 report reduced form effects of taking a second math class in 6th grade on several math outcomes during high school. These effects are estimated using the same reduced form methods as table 6 panel A—assuming no effect of treatment in 6th grade on future treatment. However, the estimation sample is smaller and the estimates often much less precise; the sample is limited to cohorts who have reached 9th (10th, 12th) grade by the 2012-13 school year, and students who remained enrolled in Miami-Dade schools through that grade. Columns 3 and 7 report the mean of each outcome among the students in the estimation sample.

There is little evidence of re-emergence in math, at least during the high school years. I find no statistically significant effects of a two-math-class treatment in 6th grade on the probability of enrolling in Algebra I (or a higher level course) during 9th grade, nor enrolling in Algebra II (or higher) by the end of high school.[25] These two benchmark outcomes are often

[25] For a subset of 6th grader cohorts in my analysis sample I also have data on course grades during 9th grade. For this even smaller sample, I estimated treatment effects on passing Algebra I by the end of 9th grade, and earning a B or higher in Algebra I by the end of 9th grade. Additionally, for one cohort I have grades for all high school years, 9th through 12th grades. With these data I estimated treatment effects on three outcomes each measured at the end

cited as indicating students who are "on track" in high school mathematics, and prepared for college mathematics. There are similarly no significant effects on math test scores at the end of 9th grade or 10th grade.[26] However, these high school differences are much less precisely estimated than outcomes in middle school.

The imprecision is an important consideration. For example, consider end-of-10th grade test scores for students at the 24th percentile of the 5th grade distribution (i.e., the level 1/level 2 cut); I cannot rule out a positive math gain of 0.25 student standard deviations, nor can I rule out a loss of 0.05 standard deviations. A 0.25 standard deviation increase in test scores would certainly be important, and change the apparent pattern of persistence. Similarly for students near the 50th percentile (the level 2/level 3 cut), I cannot rule out a 5 percentage point increase or 8 percentage point decrease in the probability of completing Algebra I by 9th grade. With time, the aging of additional cohorts should allow for tighter estimates.[27]

Columns 2 and 6 of table 8 report estimates analogous to panel B of table 6; estimates which are (partially) corrected for the dynamic nature of the treatment regime. These estimates do account for how treatment in 6th grade to effects treatment in 7th and 8th grade, but do not model any differential resources or treatments in 9th grade or beyond. Equation 5 becomes

---

of high school: having passed Algebra II, having earned a B or higher in Algebra II, and cumulative GPA in math classes. Results for these grade-related outcomes are presented in the online appendix table A2. Treatment had no statistically significant effect on these outcomes, though the reduced samples contribute to much less precision.

[26] Estimation samples for 9th and 10th grade test scores are further limited because Florida stopped giving a general end-of-grade 9th grade math test after 2009-10 and stopped giving a 10th grade test after 2010-11. The state switched to end-of-course exams.

[27] Two estimates in table 8—the effects for 9th and 10th grade tests for students at the level 2/level 3 cut-score—are precisely estimated enough to reject a naïve prediction one might have made based on the test gains measured at the end of 6th grade. Specifically, I can reject the naïve prediction formed by multiplying (i) the bivariate OLS coefficient from the regression of 9th (10th) grade score on 6th grade score, and (ii) the estimated treatment effect on 6th grade score. These predictions are 0.102 and 0.097 student standard deviations for 9th and 10th grade scores respectively. I cannot reject a similar naïve prediction for other outcomes and samples in table 8.

$$\delta^{M}_{t,t+\tau} = \delta^{T}_{t,t+\tau} - \pi_{t,t+1}\delta^{M}_{t+1,t+\tau} - \pi_{t,t+2}\delta^{T}_{t+2,t+\tau},$$

(6)

where $\tau = 3$ for outcomes in 9th grade, $\tau = 4$ for 10th grade test score, and $\tau = 7$ for outcomes by the end of high school. The estimates based on equation 6 in columns 2 and 6 are quite similar to the reduced form estimates in columns 1 and 5.

5. Effects on Achievement in Other Subjects and Attainment

Focusing attention only on math outcomes hides potentially important costs in other subjects. If the school day is fixed, allocating more time to math must reduce time spent in other subjects. As shown in figure 2, in Miami's middle schools a second, remedial math class most often crowded out a P.E. class. Other treated students missed out on classes in the arts and music or foreign languages. First, this crowd out will directly reduce achievement in the forgone subjects during the short run (treatment year), which may alter the trajectory of longer run achievement. Second, assuming math is more cognitively taxing than P.E. or arts for treated students, then the crowd out may also reduce effort in other classes, like English language arts; or reduce homework effort by increasing the marginal value of students' leisure time. To better understand these potential costs I examine treatment effects on outcomes in non-math subjects during the treatment year and during high school. However, because fewer cohorts of students have aged through high school, those analysis samples are smaller and effect estimates much less precise.

First consider effects during the treatment year. I find that taking a second, remedial math class during 6th does not change *reading* test scores at the end of 6th grade. The reading test point estimates shown in table 9 row 1 are estimated with the same sample and multi-site multi-

treatment FRD methods as the math test estimates in table 5 column 1. For students near the 50th percentile of math achievement, the estimated treatment effect is negative but very small (0.007 student standard deviations) and not statistically significant.[28]

Based on this one measure—reading achievement—increasing math instruction and crowding out some other subject does not appear to harm outcomes in non-math subjects that remain on the student's schedule during the treatment year. Yet there are many unmeasured outcomes. At a minimum, students give up the consumption value of time spent in the arts, physical activity, and other subjects. Additionally, recent research on 5th grade boys from Cawley, Frisvold, and Meyerhoefer (2012) suggests missing out on P.E. time has a causal effect on BMI and the probability of childhood obesity.

Now consider effects into the high school years, especially effects on the subjects crowded out by treatment: P.E., arts, or foreign languages in the Miami case. Achievement in these subjects is not tracked with standardized tests, but some outcome measures are available from students' high school transcripts. The estimation sample for these non-math outcomes is limited to cohorts who have reached 12th grade by the 2012-13 school year, and students who remained enrolled at Miami-Dade schools through 12th grade.

Figure 10 plots the proportion of students who completed two years (or more) of foreign language by the end of high school. Two years of foreign languages is often a (stated) requirement for admission to selective colleges and universities. No discontinuities are apparent in figure 10, but, as discussed above, this reduced form evidence may obscure effects since figure 10 does not account for other treatments or the dynamic treatment assignment. Table 9

[28] The lack of treatment effect on reading scores in 6th grade, the treatment year, is also true in the years after treatment: 7th through 10th grades. In results not presented here, I examined reading scores using the same methods developed for math scores. As the 6th grade effect, point estimates were close to zero and not statistically significant.

row 3 reports the estimated local average treatment effect, adjusted and scaled-up using the multi-site multi-treatment FRD method. In these adjusted estimates, students at the 50th percentile of the 5th grade math distribution who take a second math class during 6th grade are about 10 percentage points less likely to have completed two years of foreign languages by the end of high school. Approximately one-quarter of the analysis sample completed two years of foreign language (column 7). A 10 point effect is large, especially given the comparatively small displacement in foreign language taking during the treatment year and no displacement during 7th or 8th grade discussed in section 2. This effect should be interpreted with caution. First, it is the sole difference among several outcomes tested in tables 8 and 9. Second, the estimates are much less precise. I cannot reject essentially zero effect; the upper 95 percent confidence interval is less than a 1 point decline.[29]

By contrast, I find no evidence that treatment changed students P.E. or music and arts course taking during high school, though the outcome measures here are admittedly simple. Rows 4 and 5 of table 9 report treatment effects on the number of years a student took a course in P.E. and music or arts respectively. The point estimates are sometimes positive, sometimes negative, and small relative to the sample averages of about 1.5 P.E. and 1.5 arts classes. No difference is statistically significant, though quite imprecise. The largest potential difference is at the level 1/level 2 cut-score where I cannot rule out a reduction of up to one semester of arts or music courses.[30]

[29] A naïve prediction (as detailed in footnote 27) based on the correlation between 6th grade test scores and high school foreign language attainment would have predicted a 3 point increase in the probability of completing two years. I can reject this naïve prediction, but cannot reject any other similarly formed naïve prediction for the outcomes and samples in table 9.

[30] For one cohort of students I have data on course grades for all four high school years. For this one cohort sample, using the same methods as table 9, I estimated treatment effects on GPA excluding math classes, and on GPA for core non-math classes (i.e., ELA, science, and social studies). I also examined P.E. GPA, arts and music GPA, and foreign language GPA for students who took at least one class in the given subject. Results are presented in the online appendix table A2. Treatment had no effect on any of these course grade outcomes.

Finally, I find no effect of the two math class treatment in 6th grade on the probability that students persist through high school and graduate on time. The year following a student's first 9th grade year is a critical transition point; most high school dropouts leave at this point and many students on the margin of dropping out end up repeating 9th grade. However, row 2 of table 9 shows no treatment effect on the probability of enrolling in 10th grade the year after a student's first 9th grade year. Row 6 also shows no statistically significant effect of treatment on the probability of graduating high school on time. However, a 5 percent increase in graduation, if precisely estimated, would be an important benefit to consider in the policy calculus. And, again, I cannot rule out even larger changes in the chances of graduation; I cannot reject a 16-20 percent increase over the sample average graduation rate, nor can I reject a 6 percent decline.

6. Conclusion

This paper first provides causal evidence for a perhaps unsurprising first result: doubling the typical amount of class time devoted to math instruction substantially increases the math test scores of relatively low-achieving middle-school students. Among students quasi-randomly assigned to take two math classes—one remedial and one traditional—instead of the traditional one math class schedule, contemporaneous math scores rose by 0.16-0.18σ.

Yet, like many other educational interventions studied empirically, those initial gains did not fully persist in the school years after students returned to a regular schedule. One year after treatment ended only one-third to one-half of the initial gain remained. Two years out the effects had shrunk to one-third the original size. Once students reach high school I find little evidence of differences in math achievement or outcomes in other subjects, though the estimates are

comparatively less precise and I cannot rule out what would be meaningfully large benefits (costs).

This pattern of decaying effects in the years following treatment is similar to alternative strategies for improving achievement, like reducing class size or improving the effectiveness of teachers. That similarity suggests a need to reconsider whether current remedial education strategies—characterized by short-lived increases in the quantity of instruction—are a cost-effective way to raise the math achievement of students who currently lag expectations for their age. First, and importantly, allocating more of the school day to math imposes an opportunity cost of missed instruction in other subjects. In Miami treated students missed out on physical education, music and arts, or foreign language classes.

Increasing math instruction also carries labor costs. During the school years I study in this paper, the remediation program in Miami-Dade created roughly one "Intensive Math" class for every seven regular math classes. Put differently, the district needed 15 percent more math teachers. While some of the salary costs may be offset by reductions in teachers of other subjects, the costs of recruitment in the relatively tight math-teacher market are an important consideration. Additionally, even if a district, like Miami-Dade, is able to substantially expand their math teacher workforce without a loss of quality, the general equilibrium effect on math teacher demand must be felt somewhere, presumably in some other district. By contrast, an intervention that seeks to improve math achievement by boosting teacher-performance would require better selection, development, or job assignment of a school's existing math teachers. Selection and development, as mechanisms for improving the quality of instruction, receive the most attention from policy makers and researchers, but the evidence is mixed (Yoon et al. 2007, Taylor and Tyler 2012, Rothstein 2012).

Given the potential costs of crowding out other subjects, a natural alternative proposal would be to increase the total amount of instructional time, either with more or longer school days. The existing evidence on the effectiveness of this alternative is not clear, as discussed in section 1. Moreover, increasing total time would crowd out current out-of-school activities, and would similarly raise demand for math teachers.

In documenting the initially large but fading gains induced by doubling math instruction, the estimates presented in this paper add to a growing literature on the role of quantity of instruction in educational production. Students' seemingly mundane subject-by-subject class schedules are an important allocation of a scarce resource with potentially complex long-run effects. This decision remains an understudied area in the economics of education.

Acknowledgements

I greatly appreciate the support of the Miami-Dade County Public Schools. Susanna Loeb and Sean Reardon provided feedback throughout. I also thank Tom Dee, Eric Bettinger, Phil Gleason, Nora Gordon, the editor and referees, as well as seminar participants at Cornell University, Stanford University, the Association for Public Policy Analysis and Management, and the Association for Education Finance and Policy. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090016 to Stanford University. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or the Miami-Dade County Public Schools.

References

Angrist, Joshua D., Parag A. Parthak, and Christopher R. Walters. 2011. "Explaining Charter School Effectiveness." National Bureau of Economic Research Working Paper 17332.

Angrist, Joshua D and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics.* Princeton, NJ: Princeton University Press.

Bandeira de Mello, Victor. 2011. *Mapping State Proficiency Standards Onto the NAEP Scales: Variation and Change in State Standards for Reading and Mathematics, 2005–2009.* NCES 2011-458. Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences.

Bahrick, Harry P. and Lynda K. Hall. 1991. "Lifetime Mantinence of High School Mathematics Content." *Journal of Experimental Psychology: General,* 120 (1): 20-33.

Brown, Byron W. and Daniel H. Sacks. 1987. "The Micreconomics of Allocation of Teachers' Time and Student Learning." *Economics of Education Review* 6 (4): 319-332.

Bryk, Anthony. S., Valerie E. Lee, and Peter B. Holland. 1993. *Catholic Schools and the Common Good.* Cambridge, MA: Harvard University Press.

Carrell, Scott E. and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy,* 118 (3): 409-432.

Carroll, John. B. 1963. "A Model of School Learning." *Teachers College Record* 64: 723-733.

Cascio, Elizabeth U. and Douglas O. Staiger. 2012. "Skill, Standardized Tests, and Fadeout in Educational Interventions." National Bureau of Economic Research Working Paper 18038.

Cawley, John, David Frisvold, and Chad Meyerhoefer. 2012. "The Impact of Physical Education on Obesity Among Elementary School Children." National Bureau of Economic Research Working Paper 18341.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investment: Evidence from a Dynamic Regression Discontinuity Design." *Quarterly Journal of Economics* 125 (1): 215-261.

Chamberlain, Gary and Guido Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica* 72 (1): 295-306.

Checkoway, Amy, Beth Boulay, Beth Gamse, Meghan Caven, Lindsay Fox, Kristina Kliorys, Rachel Luck, Kenyon Maree, Melissa Velez, and Michelle Woodford. 2011. "Evaluation of the Expanded Learning Time Initiative Year Four Integrated Report: 2009-10." Cambridge, M.A.: Abt Associates Inc.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2013a. "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." National Bureau of Economic Research Working Paper 19423.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2013b. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research Working Paper 19424.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR." *Quarterly Journal of Economics* 126 (4): 1593-1660.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. "Who Teaches Whom? Race and the Distribution of Novice Teachers." *Economics of Education Review* 24 (4): 377-392.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41 (4): 778-820.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2011. "The Aftermath of Accelerating Algebra: Evidence from a District Policy Initiative." National Bureau of Economic Research Working Paper 18161.

Cooper, Harris, Kelly Charlton, Jeff C. Valentine, Laura Muhlenbruck, and Geoffrey D. Borman. 2000. "Making the Most of Summer School: A Meta-Analytic and Narrative Review." *Monographs of the Society for Research in Child Development* 65 (1): 1-127.

Cortes, Kalena, Joshua Goodman, and Takako Nomi. Forthcoming. "Intensive Math Instruction and Educational Attainment: Long-run Impacts of Double-Dose Algebra." *Journal of Human Resources.*

Dee, Thomas S. and Brian A. Jacob. 2010. "The Impact of No Child Left Behind on Students, Teachers, and Schools." *Brookings Papers on Economic Activity* Fall: 149-207.

Dougherty, Shaun. 2012. "Bridging the Discontinuity in Adolescent Literacy: Evidence of an Effective Middle Grades Intervention." Paper presented at the APPAM Fall Research Conference, November 2012.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739-1774.

Duncan, Greg J., Pamela A. Morris, and Chris Rodrigues. 2011. "Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments." *Developmental Psychology* 47 (5): 1263-1279.

Ericsson, K. Anders, Ralf Th. Krampe, and Clemens Tesch-Romer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review*, 100 (3): 363–406.

Gick, Mary L., and Keith J. Holyoak. 1983. "Schema Induction and Analogical Transfer." *Cognitive Psychology*, 15 (1): 1–38.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." *Econometrica* 69: 201–209.

Hanushek, Eric A. and Steven G. Rivkin. 2010. "Generalizations About Using Value-Added Measures of Teacher Quality." *American Economic Review* 100 (2): 267-271.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2 (3): 172-177.

Hoxby, Caroline. 2002. "The Power of Peers: How Does the Makeup of a Classroom Influence Achievement?" *Education Next* 2 (2): 57-63.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. 2009. "How New York City's Charter Schools Affect Achievement, August 2009 Report." Cambridge, MA: New York City Charter Schools Evaluation Project.

Imbens, Guido W. and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142 (2): 615-635.

Jacob, Brian A. and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86 (1): 226-244.

Jacob, Brian A., Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45 (4): 915-943.

Jacob, Brian A. and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics,* 188 (3): 843-877.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14601.

Kemple, James J., Corinne Herlihy, and Thomas J. Smith. 2005. "Making Progress Toward Graduation: Evidence from the Talent Development High School Model." New York, NY: MDRC.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz., 2007. "Experimental Analysis of Neighborhood Effects" *Econometrica* 75 (1): 83-119.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497-532.

Krueger, Alan B and Diane M Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *Economic Journal* 111 (468): 1-28.

Lavy, Victor. 2010. "Do Differences in School's Instruction Time Explain International Achievement Gaps in Math, Science, and Reading? Evidence from Developed and Developing Countries." National Bureau of Economic Research Working Paper 16227.

Lavy, Victor and Analia Schlosser. 2005. "Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits." *Journal of Labor Economics* 23 (4): 839-874.

Lazear, Edward P. 2001. "Educational Production." *Quarterly Journal of Economics* 116 (3): 777-803.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142 (2): 655-674.

Loeb, Susanna, Demetra Kalogrides, and Tara Béteille. Forthcoming. "Effective Schools: Teacher Hiring, Assignment, Development, and Retention." *Education Finance and Policy.*

Ludwig, Jens. and Douglas L. Miller. 2007. “Does Head Start Improve Children’s Life Chances? Evidence From a Regression Discontinuity Design.” *Quarterly Journal of Economics* 122 (1): 159–208.

McCrary, Justin. 2008. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics* 142 (2): 698-714.

National Center for Education Statistics. 2011. *The Nation’s Report Card: Mathematics 2011.* NCES 2012-458. Institute for Education Sciences, U.S. Department of Education, Washington D.C.

Nomi, Takako and Elaine Allensworth. 2009. “’Double-Dose’ Algebra as an Alternative Strategy to Remediation: Effects on Students' Academic Outcomes.” *Journal of Research on Educational Effectiveness* 2 (2): 111-148.

Nomi, Takako and Stephen W. Raudenbush. 2013. “Academic Differentiation, Classroom Peer Skill, and Inequality: Evidence from a Natural Experiment in 60 Urban High Schools.” Working paper, February 17, 2013.

Patall, Ericka A., Harris Cooper, and Ashley Batts Allen. Forthcoming. “Extending the School Day or the School Year: A Systematic Review of Research (1985-2009).” *Review of Educational Research.*

Raudenbush, Stephen W., Guanglei Hong, and Brian Rowan. 2002. “Studying the Causal Effects of Instruction with Application to Primary-School Mathematics.” Paper presented at Research Seminar II: Instructional and Performance Consequences of High-poverty Schooling, National Center for Education Statistics, March 11.

Raudenbush, Stephen W., Sean F. Reardon, and Takako Nomi. 2012. “Statistical Analysis for Multisite Trials Using Instrumental Variables with Random Coefficients.” *Journal of Research on Educational Effectiveness,* 5 (3): 303-332.

Reardon, Sean F. and Stephen W. Raudenbush. 2013. “Under What Assumptions do Site-by-Treatment Instruments Identify Average Causal Effects?” *Sociological Methods and Research* 42 (2): 143-163.

Reardon, Sean F., Fatih Unlu, Pei Zhu, and Howard S. Bloom. 2014. “Bias and Bias Correction in Multisite Instrumental Variables Analysis of Heterogeneous Mediator Effects.” *Journal of Educational and Behavioral Statistics* 39 (1): 53-86.

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” *Quarterly Journal of Economics* 25 (1): 175-214.

Rothstein, Jesse. 2012. “Teacher Quality Policy When Supply Matters.” National Bureau of Economic Research Working Paper 18419.

Schwerdt, Guido and Martin R West. 2012. “The Effects of Early Grade Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida.” Harvard University, Program on Education Policy and Governance Working Paper 12-09.

Sims, David P. 2008. “Strategic Responses to School Accountability Measures: It's All In the Timing.” *Economics of Education Review* 27 (1): 58-68.

Taylor, Eric S. and John H. Tyler. 2012 “The Effect of Evaluation on Teacher Performance” *American Economic Review* 102 (7):

Zepeda, Sally J. and R. Stewart Mayers. 2006. “An Analysis of Research on Block Scheduling.” *Review of Educational Research* 76 (1): 137-170.

Yoon , Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement: Issues & Answers Report, REL 2007–No. 033*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

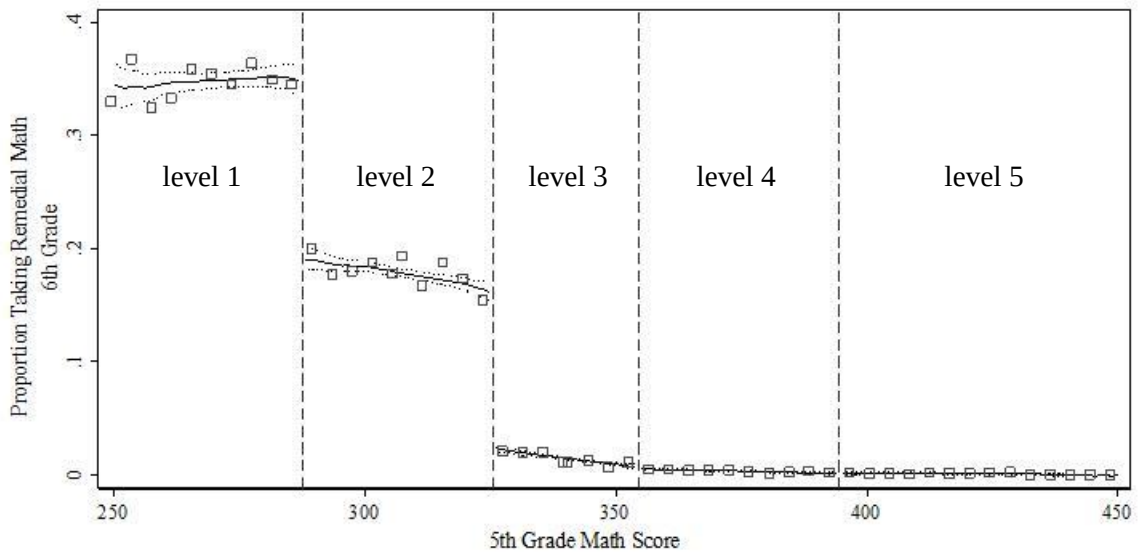


FIGURE 1—PROPORTION STUDENTS TAKING A SECOND, REMEDIAL MATH CLASS IN 6TH GRADE BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean of an indicator = 1 if the student took two math classes in 6th grade (the treatment of interest) (y-axis), within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

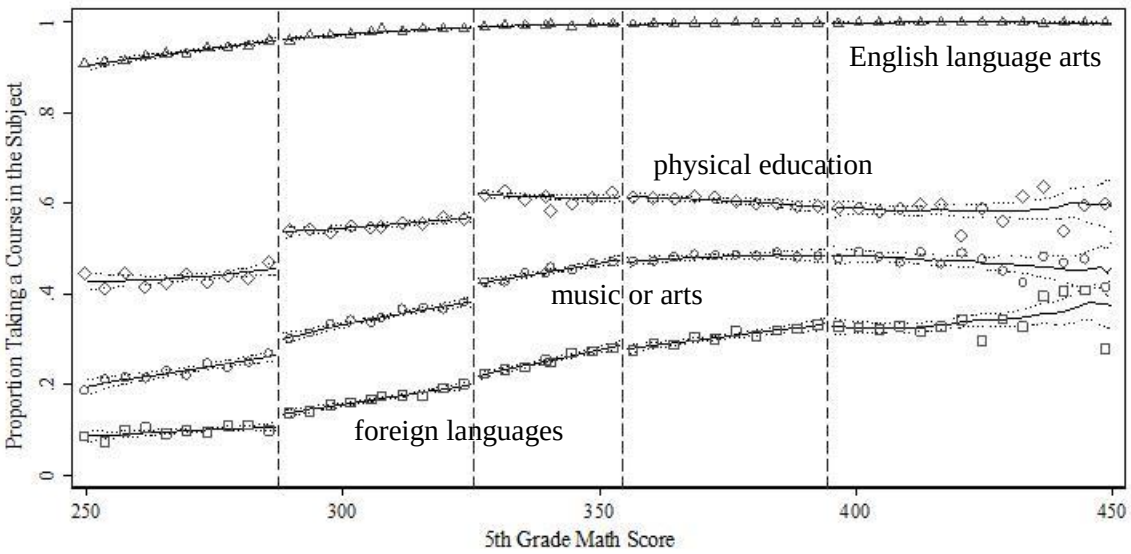


FIGURE 2—PROPORTION STUDENTS TAKING A CLASS IN VARIOUS SUBJECT AREAS IN 6TH GRADE BY 5TH GRADE TEST SCORE

NOTE: Markers represent the mean of an indicator = 1 if the student took a class in foreign languages (squares), music or arts (circles), physical education (diamonds), and English language arts (triangles) in 6th grade (y-axis), within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

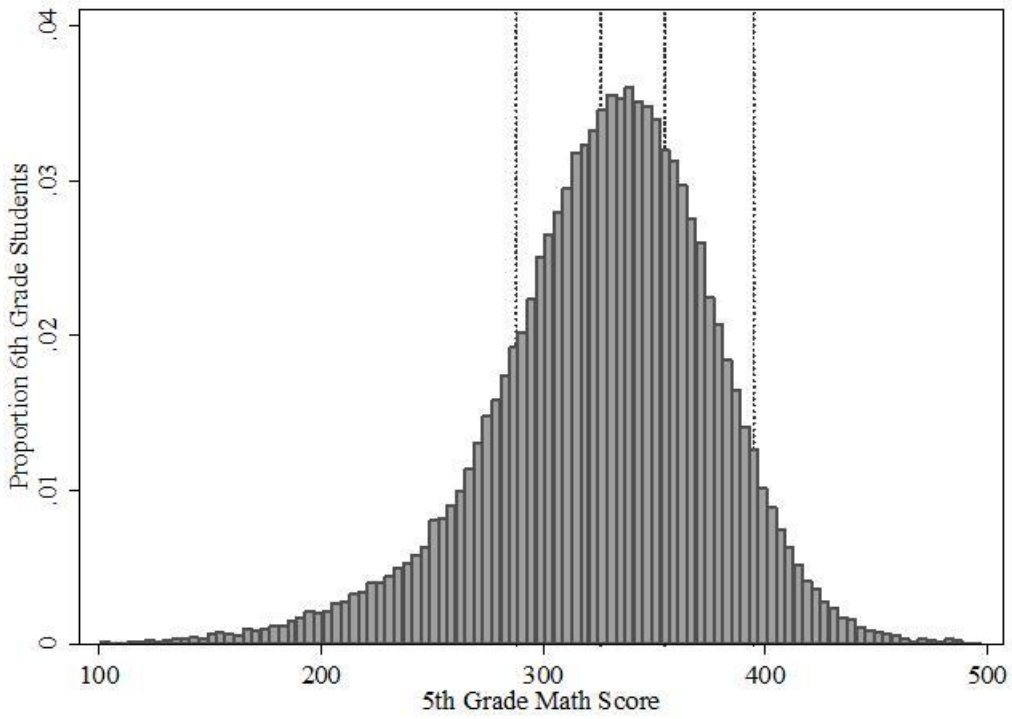


FIGURE 3—DISTRIBUTION OF 5TH GRADE MATH TEST SCORES (FORCING VARIABLE)

NOTE: Bars measure the proportion of 6th grade students receiving at a particular scale score on the 5th grade end of grade math test, in bins of four scale score points. Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test.

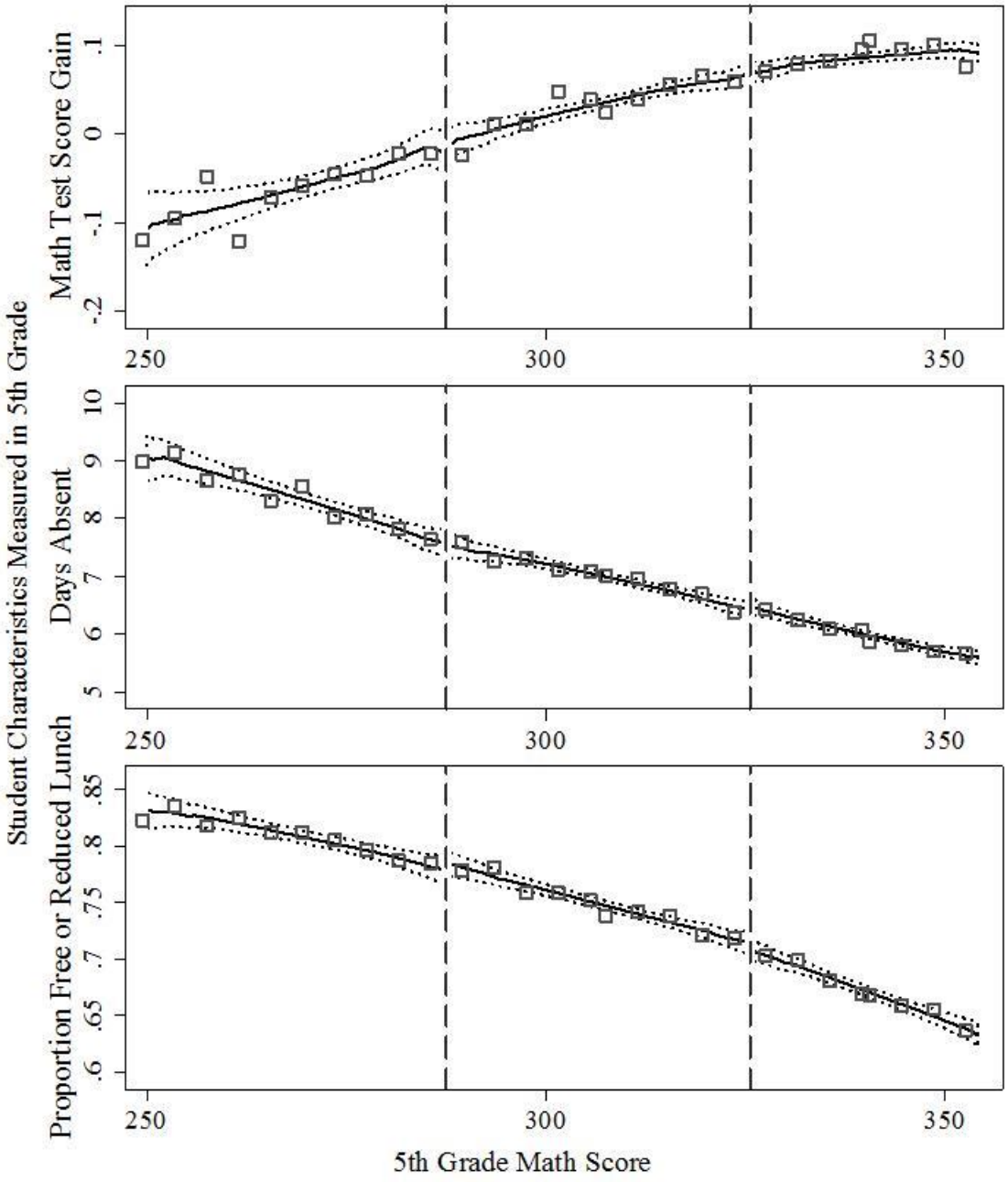


FIGURE 4—PRE-TREATMENT STUDENT CHARACTERISTICS BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean of the characteristic described on the y-axis, within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

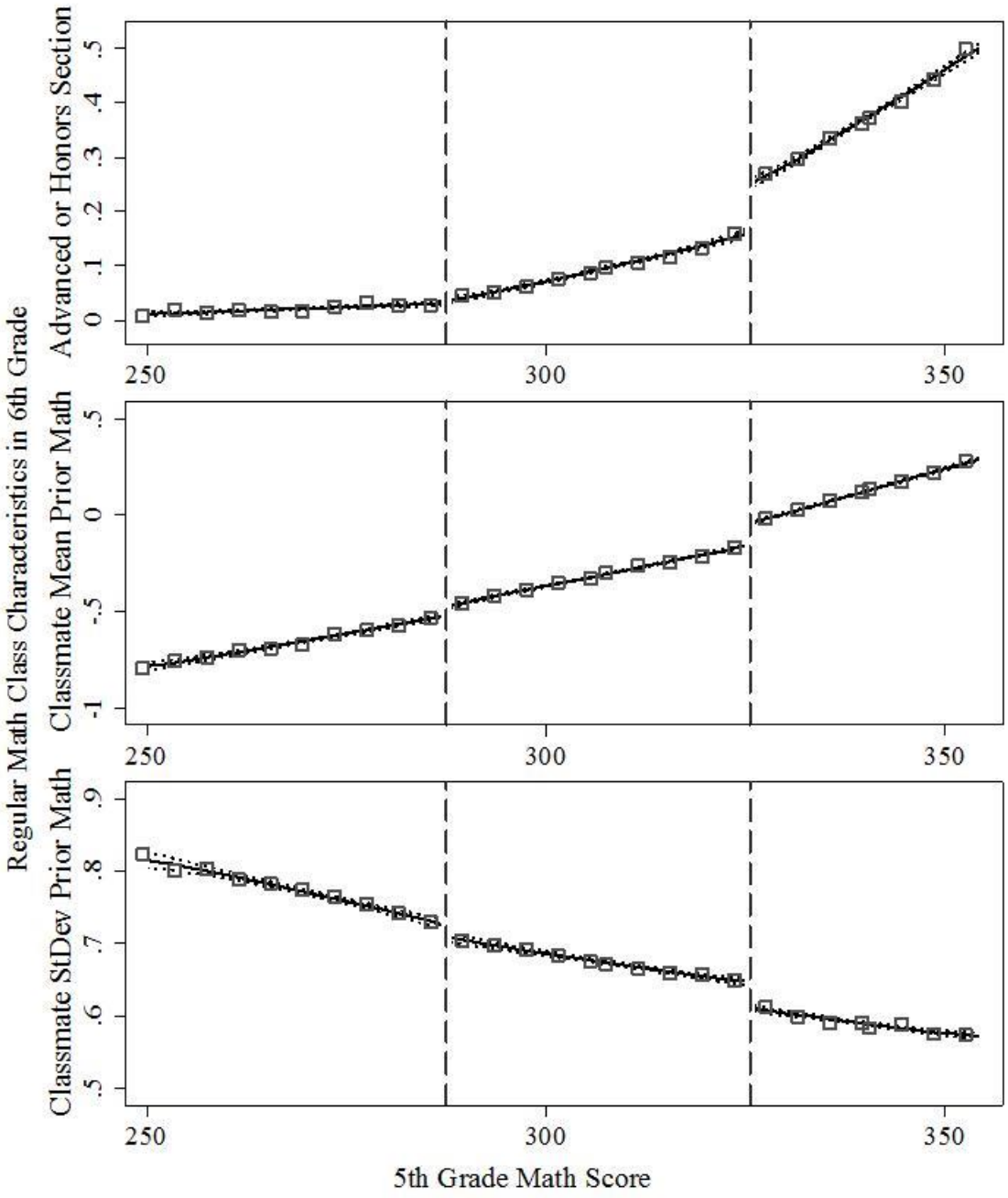


FIGURE 5— CHARACTERISTICS OF REGULAR MATH CLASS IN 6TH GRADE BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean of the characteristic described on the y-axis, within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

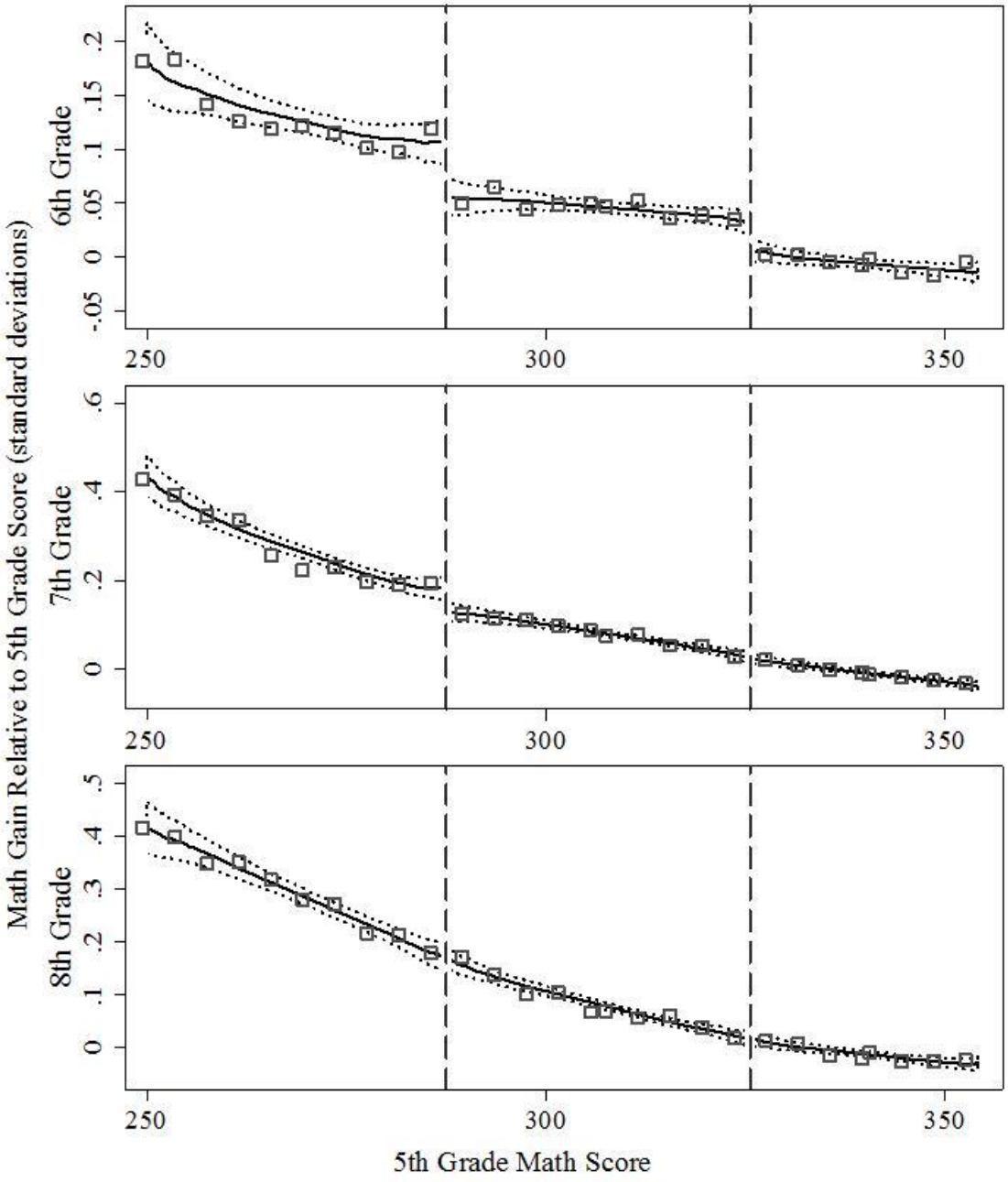
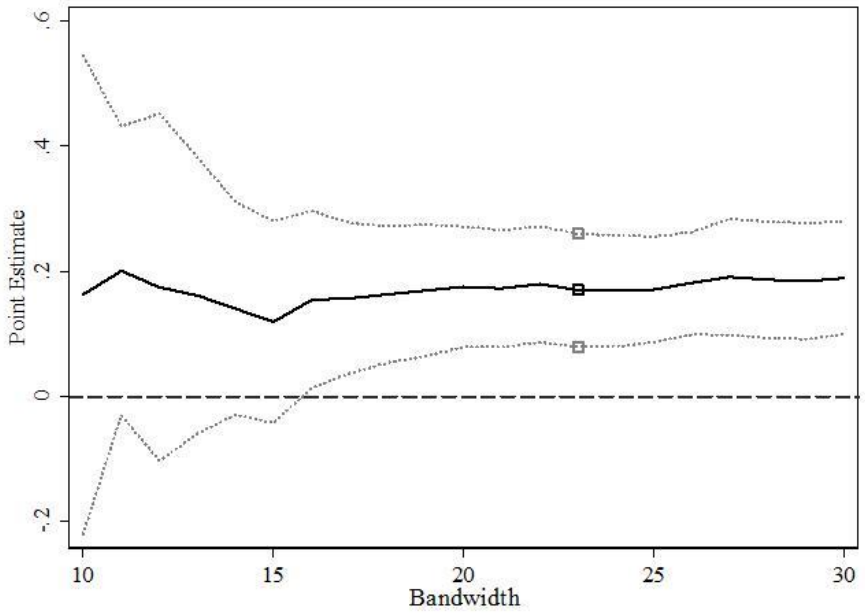


FIGURE 6—STUDENT MATH ACHIEVEMENT GAINS OVER TIME
BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean math test score gain 6th grade score minus 5th grade score (top panel), 7th minus 5th gain (middle panel), and 8th minus 5th gain (bottom panel) (y-axis), within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

PANEL A—LEVEL 1/LEVEL 2 CUT-SCORE



PANEL B—LEVEL 2/LEVEL 3 CUT-SCORE

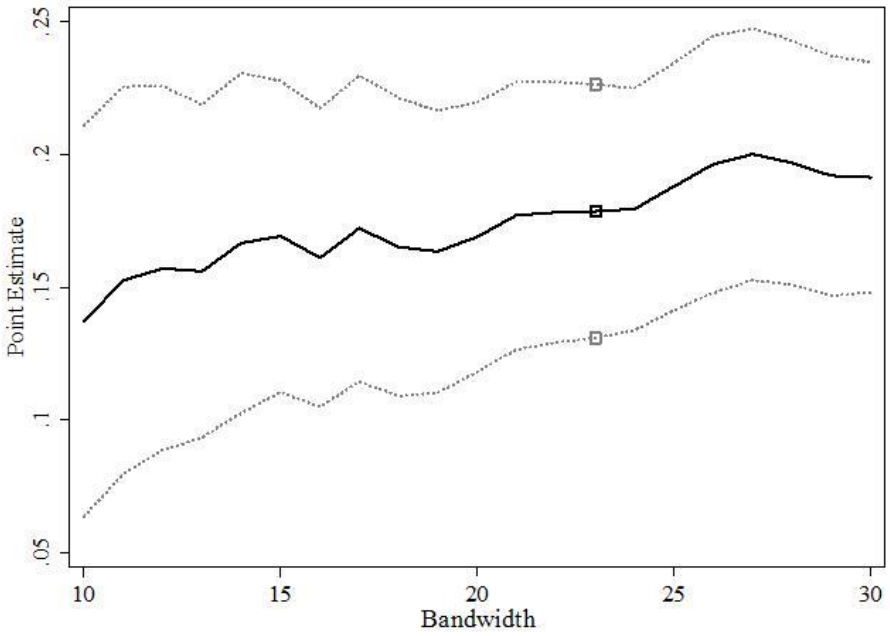


FIGURE 7— ESTIMATED TREATMENT EFFECT ON MATH TEST SCORES AT THE END OF 6TH GRADE BY VARYING BANDWIDTHS

NOTE: The solid line traces out a series of treatment effect point estimates (y-axis). Each point is estimated with the same multi-site FRD methods used for table 5 columns 1 and 5, except that each point is estimated with a different integer bandwidth (x-axis). The dotted lines trace out the 95 percent confidence intervals. The square markers indicate the point estimate corresponding to the optimal bandwidth 23 reported in table 5, row 1, columns 1 and 5.

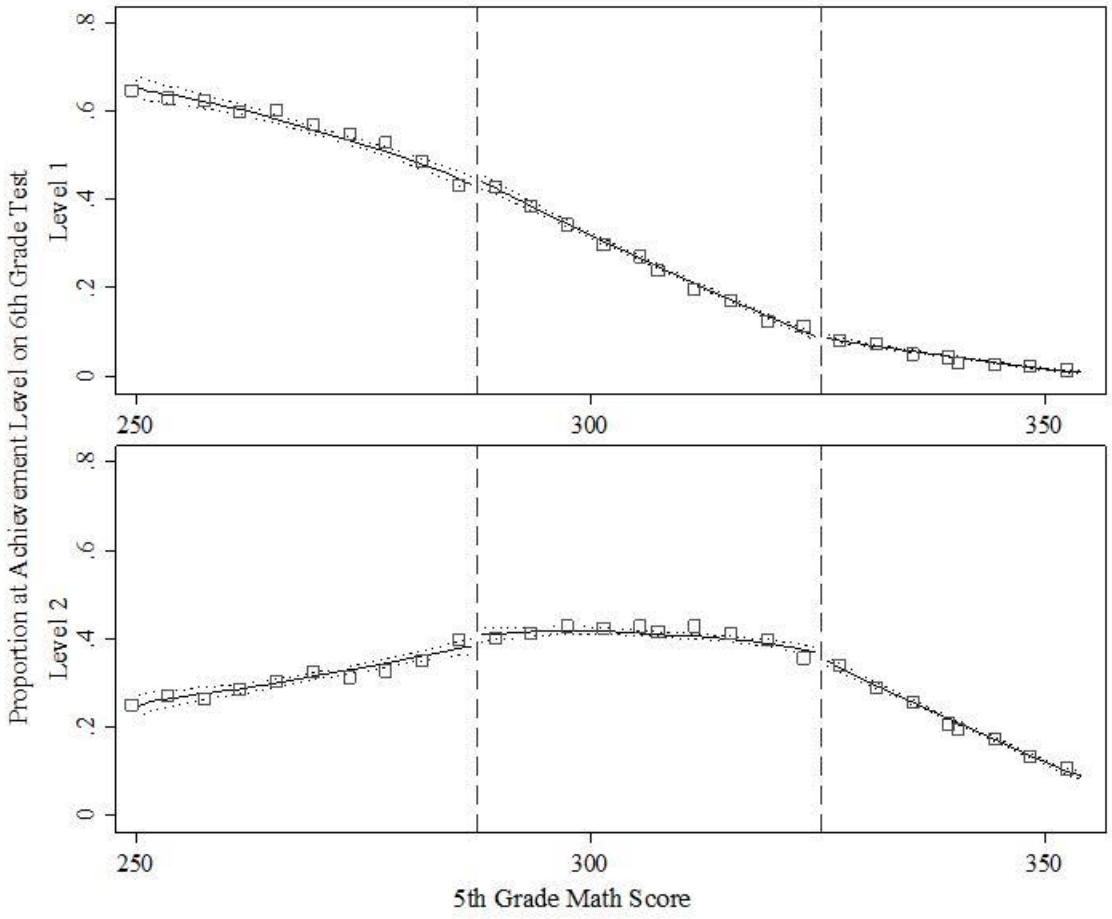
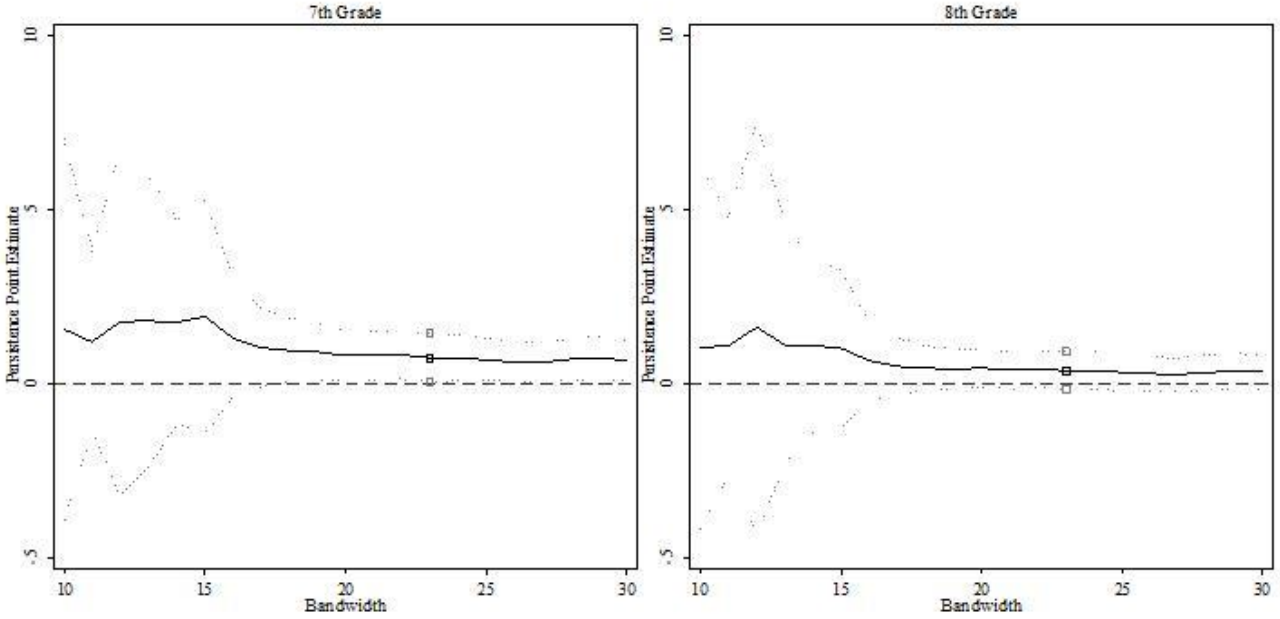


FIGURE 8—PROPORTION STUDENTS SCORING AT ACHIEVEMENT LEVEL 1 OR 2 ON THE 6TH GRADE MATH TEST BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean of an indicator = 1 if the student's 6th grade math test score placed them in "achievement level 1" (top panel), or in "achievement level 2" (bottom panel) (y-axis), within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval.

PANEL A—LEVEL 1/LEVEL 2 CUT-SCORE



PANEL B—LEVEL 2/LEVEL 3 CUT-SCORE

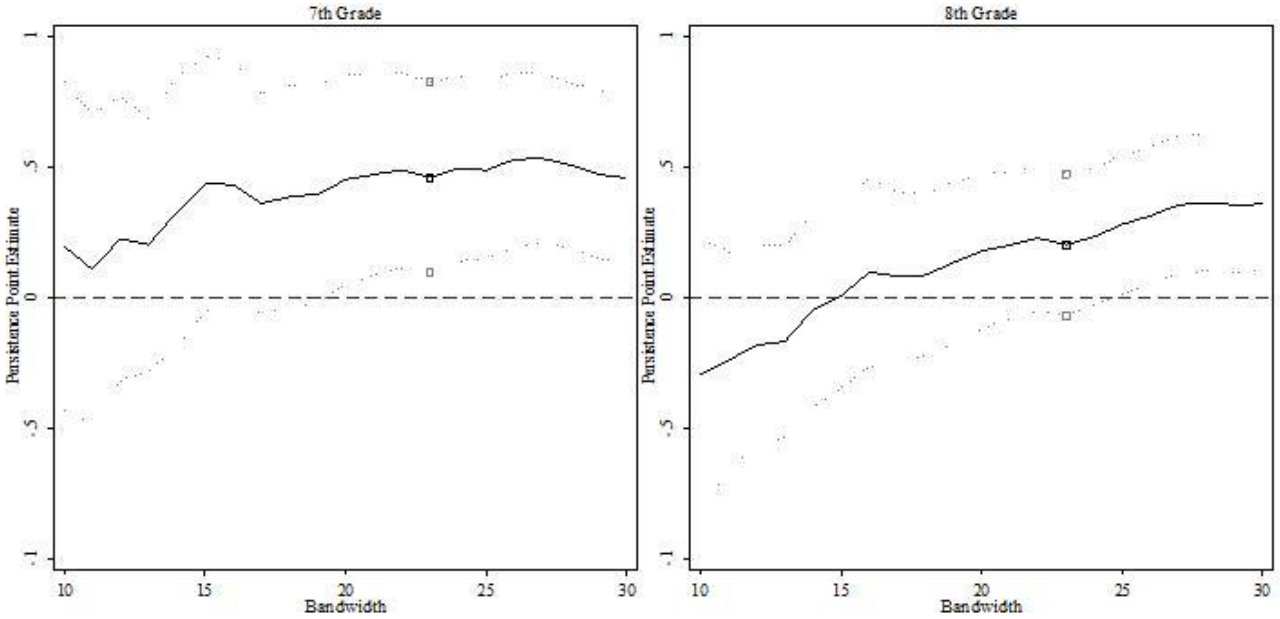


FIGURE 9— ESTIMATED PERSISTENCE OF GAINS IN 6TH GRADE AT THE END OF 7TH AND 8TH GRADE BY VARYING BANDWIDTHS

NOTE: The solid line traces out a series of persistence point estimates (y-axis). Each point is estimated with the same dynamic-treatment-assignment multi-site FRD methods used for table 6 panel B, except that each point is estimated with a different integer bandwidth (x-axis). The dotted lines trace out the 95 percent confidence intervals. The square markers indicate the point estimate corresponding to the optimal bandwidth 23 reported in table 6, panel B, row 2.

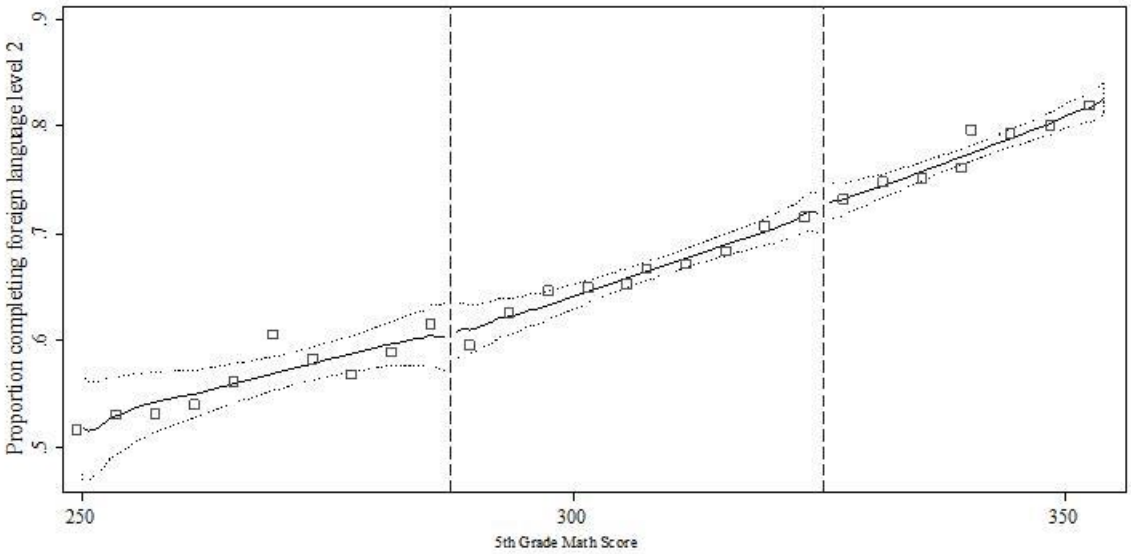


FIGURE 10—PROPORTION STUDENTS COMPLETING TWO YEARS OF FOREIGN LANGUAGES AT THE END OF HIGH SCHOOL BY 5TH GRADE TEST SCORE

NOTE: Square markers represent the mean of an indicator = 1 if the student completed two years of foreign language by the end of high school, (y-axis), within bins of four scale score points on the 5th grade math test (x-axis). Vertical dashed lines mark the cut-scores dividing "achievement levels" on the 5th grade test. Local linear fitted lines (solid lines) are estimated on all student-level data, within achievement levels, using a rectangular kernel and a bandwidth of 23 scale score points. Dotted lines trace the 95 percent confidence interval. Sample limited to cohorts who began 6th grade in 2004-05 through 2007-09, and who remained enrolled in the district through 12th grade.

Table 1—Course taking during the treatment year

	5th grade math test score below…			
	level 1/2 cut-score	level 2/3 cut-score	level 3/4 cut-score	level 4/5 cut-score
	(1)	(2)	(3)	(4)
Took a second math class	0.1889**	0.1271**	0.0024	-0.0001
	(0.0117)	(0.0070)	(0.0019)	(0.0017)
Took a class in …				
Physical education	-0.0696**	-0.0444**	-0.0013	0.0254
	(0.0145)	(0.0070)	(0.0091)	(0.0212)
Music or arts	-0.0423**	-0.0377**	0.0017	0.0038
	(0.0119)	(0.0092)	(0.0109)	(0.0144)
Foreign language	-0.0300*	-0.0182*	0.0157	-0.0051
	(0.0119)	(0.0070)	(0.0102)	(0.0170)
English language arts	0.0006	-0.0001	-0.0000	0.0001
	(0.0004)	(0.0002)	(0.0002)	(0.0001)
Science	-0.0007	0.0003	0.0001	a
	(0.0006)	(0.0004)	(0.0003)	
Social studies	0.0011	-0.0001	0.0001	-0.0000
	(0.0008)	(0.0005)	(0.0004)	(0.0001)
Student Observations	18868	35637	34906	14664

Note: Each cell reports a local average treatment effect from a separate regression, estimated using standard sharp regression discontinuity methods. Each dependent variable is an indicator = 1 if, for row 1, the student took a second math course; and, for rows 2-7, if the student took a any course in the subject during their 6th grade year. The independent variables include (i) an indicator = 1 if the student scored below the given cut-score (the reported coefficient), (ii) a linear term for the forcing variable (5th grade test score), and (iii) an interaction of the indicator and forcing variable which allows the slope to differ above and below the cut-score. Estimation is by local-linear least squares using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. The estimation sample is restricted to 6th grade students who were subsequently observed in 7th and 8th grades. Standard errors are clustered at the discrete values of the forcing variable.

[a] All students in this cell took a science course.

* indicates $p<0.05$, ** $p<0.01$

Table 2—Pre-treatment student characteristics

	A: Mean (standard deviation)		*B:* Relative likelihood that compliers have the characteristic		*C:* Estimated pre-treatment discontinuity (standard error)	
	All students	Students with data for grades 6, 7, 8	level 1/2 cut-score	level 2/3 cut-score	level 1/2 cut-score	level 2/3 cut-score
	(1)	(2)	(3)	(4)	(5)	(6)
Student observations	131,172	90,262	18865	35637	18865	35637
Math test score, 5th grade	-0.019	0.148				
	(1.002)	(0.899)				
Math test score gain,	0.033	0.035			0.032	-0.003
4th to 5th grade	(0.628)	(0.587)			(0.020)	(0.010)
Reading test score gain	0.037	0.036			0.032	-0.014
4th to 5th grade	(0.641)	(0.607)			(0.017)	(0.012)
Female	0.485	0.518	1.031	1.116	0.008	0.012
					(0.014)	(0.009)
Hispanic	0.617	0.630	1.205	0.967	-0.001	0.002
					(0.014)	(0.010)
African-American	0.264	0.249	0.712	1.113	-0.004	-0.004
					(0.014)	(0.007)
White	0.093	0.095	0.884	0.883	0.010	0.005
					(0.005)	(0.008)
Free or reduced lunch	0.685	0.645	0.987	0.997	-0.012	0.009
					(0.013)	(0.007)
English language learner	0.526	0.536	1.114	1.029	0.000	0.000
					(0.013)	(0.006)
Frequently absent	0.148	0.125	1.131	0.941	0.002	-0.003
(>12 absences 5th grade)					(0.009)	(0.004)
Repeating 6th grade	0.021	0.019	1.802	0.921	0.002	0.001
					(0.004)	(0.002)

Note: Column group A: Author's calculations of means (standard deviations) for 6th grade Miami-Dade middle school students pooling the 2004-05 to 2008-09 school years. Column 2 reports statistics for the subsample of 6th grade students with complete data for 6th, 7th, and 8th grades.

Column group B: Ratio of (i) the fuzzy regression discontinuity first-stage compliance estimate (as in table 1 column 1) for students with the specified characteristic (e.g., Hispanic students), over (ii) the compliance estimate among all students. See text and Angrist and Pischke (2009, pp. 171-172) for interpretation of this ratio.

Column group C: Each cell reports a treatment "effect" on a pre-treatment variable from a separate regression, estimated using standard sharp regression discontinuity methods. Each dependent variable is the pre-treatment characteristic. The independent variables include (i) an indicator = 1 if the student scored below the given cut-score (the reported coefficient), (ii) a linear term for the forcing variable (5th grade test score), and (iii) an interaction of the indicator and forcing variable which allows the slope to differ above and below the cut-score. Estimation is by local-linear least squares using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. Standard errors are clustered at the discrete values of the forcing variable.

* indicates $p<0.05$, ** $p<0.01$ for columns 5 and 6.

Table 3—Discontinuities in the characteristics of students' regular math and English language arts classes in 6th grade

	5th grade math test score below…			
	level 1/2 cut-score	level 2/3 cut-score	level 3/4 cut-score	level 4/5 cut-score
	(1)	(2)	(3)	(4)
Regular math class characteristics				
Advanced or honors section	-0.007	-0.088**	-0.063**	-0.013
	(0.005)	(0.012)	(0.009)	(0.010)
Class mean prior math score	-0.040**	-0.104**	-0.058**	-0.029*
	(0.012)	(0.010)	(0.007)	(0.011)
Class st. dev. prior math score	0.012**	0.024**	0.012**	-0.002
	(0.004)	(0.004)	(0.003)	(0.005)
Teacher value-added measure	0.001	-0.004	-0.004	-0.001
	(0.004)	(0.002)	(0.003)	(0.003)
Teacher has master's degree	-0.023	0.006	-0.011	-0.017
	(0.016)	(0.009)	(0.008)	(0.020)
Teacher years of experience (in district)	0.235	-0.244	-0.084	-0.395
	(0.276)	(0.192)	(0.217)	(0.250)
English language arts class characteristics				
Class mean prior math score	0.003	-0.015	-0.007	-0.031
	(0.018)	(0.010)	(0.008)	(0.015)
Class st. dev. prior math score	0.002	0.002	-0.008*	-0.001
	(0.006)	(0.003)	(0.003)	(0.005)
Class mean prior ELA score	0.000	-0.022	-0.005	-0.021
	(0.024)	(0.012)	(0.009)	(0.017)
Class st. dev. prior ELA score	0.002	0.002	-0.008	0.003
	(0.006)	(0.003)	(0.004)	(0.004)
Student observations	18865	35637	34905	14664

Note: Each cell reports a local average treatment effect from a separate regression, estimated using standard sharp regression discontinuity methods. Dependent variables are the characteristics listed for each row. All class mean and standard deviation dependent variables are jackknife, excluding the student herself from the calculation. See footnote X for a description of the teacher value-added scores. The independent variables include (i) an indicator = 1 if the student scored below the cut-score given in the column header (the reported coefficient), (ii) a linear term for the forcing variable (5th grade test score), and (iii) an interaction of the indicator and forcing variable which allows the slope to differ above and below the cut-score. Estimation is by local-linear least squares using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. The estimation sample is restricted to 6th grade students who were subsequently observed in 7th and 8th grades. Standard errors are clustered at the discrete values of the forcing variable.
* indicates $p<0.05$, ** $p<0.01$

Table 4—Between school variation in assignment of treatments at the cut-scores

	Took second math class		Regular math class characteristics: Advanced or honors section		Class mean prior math score		Class st. dev. prior math score	
	level 1/2 cut	level 2/3 cut	level 1/2 cut	level 2/3 cut	level 1/2 cut	level 2/3 cut	level 1/2 cut	level 2/3 cut
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Between-school distribution of treatment discontinuity estimates								
90th percentile	0.577	0.430	0.053	0.040	0.085	0.062	0.103	0.095
75th percentile	0.341	0.241	0.015	0.000	0.050	0.014	0.056	0.054
Mean	0.196	0.133	-0.005	-0.081	-0.052	-0.105	0.015	0.026
25th percentile	0.000	0.000	-0.030	-0.123	-0.112	-0.167	-0.024	-0.011
10th percentile	-0.018	0.000	-0.071	-0.240	-0.328	-0.365	-0.064	-0.060
Correlation between level 1/2 and level 2/3 discontinuities on same treatment	-0.39		0.08		-0.01		0.03	
Correlation between different treatment discontinuities at same cut-score								
Second math class	1	1						
Advanced or honors	-0.15	-0.15	1	1				
Class mean	-0.14	-0.07	0.14	0.74	1	1		
Class st. dev.	-0.01	0.07	0.02	-0.53	-0.47	-0.77	1	1

Note: Summary statistics of school-level estimates of discontinuities at the 5th grade math test cut-scores for 145 schools. Each school-level estimate is estimated in a separate regression using the specification described in the notes for tables 1 and 3, and with the sample restricted to only observations from the given school. In all cases the estimation sample is restricted to 6th grade students who were subsequently observed in 7th and 8th grades. Distribution parameters and correlations are weighted by the number of student observations for the school.

Table 5—Effect of second, remedial math class on achievement at end of treatment year

	Prior grade math test score below …							
	level 1/2 cut-score				level 2/3 cut-score			
	Multi-site FRD	Standard FRD	Intent-to-treat RD	Student obs.	Multi-site FRD	Standard FRD	Intent-to-treat RD	Student obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A: Effect of treatment in 6th grade on grade 6 test score								
Observed in 6th, 7th, and 8th grades	0.166	0.260	0.049	18865	0.176	0.220	0.028	35637
	(0.044)	(0.099)	(0.019)		(0.024)	(0.090)	(0.012)	
	[104.5]	[261.1]			[5091.2]	[329.5]		
Observed in 6th and 7th grades	0.158	0.253	0.049	24337	0.180	0.181	0.025	45279
	(0.038)	(0.083)	(0.016)		(0.020)	(0.068)	(0.010)	
	[35.8]	[294.3]			[7265.2]	[340.9]		
Observed in 6th grade	0.143	0.301	0.051	36499	0.188	0.204	0.028	62108
	(0.043)	(0.058)	(0.010)		(0.015)	(0.057)	(0.008)	
	[37.7]	[322.2]			[3342.0]	[533.5]		
B: Effect of treatment in 7th grade on grade 7 test score								
Observed in 7th grade	0.187	0.171	0.039	38994	0.172	0.196	0.035	50887
	(0.030)	(0.032)	(0.007)		(0.017)	(0.033)	(0.006)	
	[62.0]	[989.0]			[2978.6]	[1516.0]		
C: Effect of treatment in 8th grade on grade 8 test score								
Observed in 8th grade	0.130	0.131	0.030	37463	0.130	0.181	0.032	55954
	(0.024)	(0.042)	(0.010)		(0.017)	(0.034)	(0.006)	
	[51.5]	[1416.1]			[1627.9]	[1404.5]		

Note: Each cell reports a local average treatment effect from a separate regression, estimated using various regression discontinuity methods (indicated in column headings). Treatment period and estimations samples are described in row titles. In all cases the dependent variable is standardized math test score at the end of the treatment school year. Standard errors are clustered at the discrete values of the forcing variable. F-statistic for test of joint significance of excluded instrument(s) shown in brackets.

"Intent-to-treat RD" estimates use standard sharp regression discontinuity methods. The independent variables include (i) an indicator = 1 if the student scored below the given cut-score (the reported coefficient), (ii) a linear term for the forcing variable (prior year test score), and (iii) an interaction of the indicator and forcing variable which allows the slope to differ above and below the cut-score. Estimation is by local-linear least squares using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score.

"Standard RD" estimates use standard fuzzy regression discontinuity methods. The specification is identical to the "Intent to treat RD" estimates except that it includes a treatment indicator = 1 if the student took a second math class (the reported coefficient), and the indicator for scoring below the cut-score is the excluded instrument for the treatment indicator. Estimation is by two-stage least squares.

"Multi-site FRD" estimates are also instrumental variables estimates. The specification is identical to the "Single-instrument RD" estimates except that it includes four endogenous treatment variables: (i) an indicator = 1 if the student took a second math class (the reported coefficient), (ii) an indicator = 1 if the student's regular math class was an advanced or honors section, (iii) the jackknife mean of baseline math test score among the student's regular math class peers, and (iv) the jackknife standard deviation of baseline math score among the same peers. The excluded instruments are a vector of indicator variables formed by the interaction of (i) the standard indicator variable = 1 if the student scored below the cut-score, and (ii) school-specific indicators. The specification also includes school fixed effects. Estimation is by limited information maximum likelihood.

All estimates are different from zero at $p<0.01$

Table 6—Persistence over-time of achievement gains from second, remedial math class in 6th grade

	5th grade math test score below …					
	level 1/2 cut-score			level 2/3 cut-score		
	Grade 6 test score	Grade 7 test score	Grade 8 test score	Grade 6 test score	Grade 7 test score	Grade 8 test score
	(1)	(2)	(3)	(4)	(5)	(6)
A: Total effect of treatment in 6th grade (reduced form estimate)						
Treatment effect	0.166	0.139	0.066	0.176	0.099	0.056
	(0.044)	(0.046)	(0.042)	(0.024)	(0.030)	(0.023)
Persistence of total effect in 6th grade		0.836	0.400		0.564	0.316
		(0.357)	(0.272)		(0.187)	(0.138)
B: Marginal effect of treatment in 6th grade (dynamic-treatment, indirect FRD estimate)						
Treatment effect		0.122	0.064		0.077	0.031
		(0.047)	(0.042)		(0.030)	(0.024)
Persistence of total effect in 6th grade		0.737	0.387		0.438	0.178
		(0.343)	(0.273)		(0.182)	(0.136)

Note: Panel A: Each cell in row 1 reports a local average treatment effect from a separate regression, estimated using multi-site FRD methods. The dependent variable is standardized math test score at the end of 6th grade (columns 1, 4), 7th grade (columns 2, 5), or 8th grade (columns 3, 6). The specification includes four endogenous treatment variables: (i) an indicator = 1 if the student took a second math class in 6th grade (the reported coefficient), (ii) an indicator = 1 if the student's regular math class in 6th grade was an advanced or honors section, (iii) the jackknife mean of baseline math test score among the student's regular math class peers in 6th grade, and (iv) the jackknife standard deviation of baseline math score among the same peers. The excluded instruments are a vector of indicator variables formed by the interaction of (i) the standard indicator variable = 1 if the student scored below the cut-score on the 5th grade math test, and (ii) school-specific indicators. Additional included regressors include: (i) a linear term for the forcing variable (5th grade test score), (ii) an interaction of the indicator for scoring below the cut-score and forcing variable, and (iii) school fixed effects. Estimation is by limited information maximum likelihood using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. The estimation sample is restricted to students who were observed in 6th, 7th, and 8th grades. Standard errors are clustered at the discrete values of the forcing variable. Row 2 reports persistence estimates: the ratio of the estimated 7th grade effect over the estimated 6th grade effect (columns 2, 5), and 8th grade over 6th grade (columns 3, 6).

Panel B: Row 1 reports the LATE estimates using the "indirect FRD" method described in detail in the text which explicitly models the dynamic nature of treatment assignment. As described in equation 4, the marginal effect of treatment in 6th grade on 7th grade test scores (columns 2, 5) can be written as (a) the total effect of treatment in 6th grade on 7th grade test scores (shown in panel A), minus the product of two terms (b) the effect of treatment in 6th grade on the probability of treatment in 7th grade, and (c) the effect of treatment in 7th grade on 7th grade test scores. Each of these constituent terms (a), (b), and (c) are estimated using the multi-site FRD methods described for panel A. Standard errors are calculated using the delta method. To obtain the complete covariance matrix, the various constituent terms are estimated simultaneously using the "stacked equation" approach described in the text. An analogous approach is used for the marginal effect of treatment in 6th grade on 8th grade test scores, which is detailed in equation 5 in the text. Row 2 again reports the persistence estimate ratios.

Table 7—Alternative persistence estimates

	5th grade math test score below …			
	level 1/2 cut-score		level 2/3 cut-score	
	Grade 7 test score	Grade 8 test score	Grade 7 test score	Grade 8 test score
	(1)	(2)	(3)	(4)
Prefered estimate (table 6 row 4)	0.737	0.387	0.438	0.178
	(0.343)	(0.273)	(0.182)	(0.136)
Students observed in grades 6-7 but not necessarily in grade 8	0.891		0.721	
	(0.351)		(0.223)	
Standard single-treatment FRD	0.449	0.146	0.175	-0.040
	(0.596)	(0.468)	(0.293)	(0.300)

Note: Row 1 repeated from table 6 row 4. See estimation notes in table 6. Row 2 estimation methods are identical to row 1 except that the sample is expanded to include students who have complete data in grades 6 and 7, but not necessarily in grade 8. Row 3 returns to the same estimation sample as row 1, but uses a standard fuzzy regression discontinuity approach with one endogenous treatment variable, the two math class treatment indicator. See estimation notes in table 5 for a description.

Table 8—Effect of second, remedial math class on math outcomes high school

	5th grade math test score below …							
	level 1/2 cut-score				level 2/3 cut-score			
	Total effect (reduced form)	Marginal effect (indirect FRD)	Estimation sample mean (sd)	Student obs.	Total effect (reduced form)	Marginal effect (indirect FRD)	Estimation sample mean (sd)	Student obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math test score at the end of								
Grade 9	0.0225	0.0158	-0.3569	10389	-0.007	-0.024	0.127	19499
	(0.064)	(0.064)	(0.616)		(0.051)	(0.051)	(0.521)	
Grade 10	0.099	0.102	-0.357	9194	-0.013	-0.022	0.089	18024
	(0.078)	(0.078)	(0.625)		(0.056)	(0.056)	(0.520)	
By the end of 9th grade								
Completed Algebra I	0.003	0.003	0.979	17898	-0.015	-0.016	0.991	33818
	(0.009)	(0.009)			(0.033)	(0.033)		
By the end of high school								
Completed Algebra I	-0.004	-0.004	0.999	8643	0.000	0.001	0.999	17285
	(0.004)	(0.004)			(0.005)	(0.006)		
Completed Algebra II	-0.021	-0.024	0.898	8643	-0.005	-0.005	0.939	17285
	(0.037)	(0.037)			(0.023)	(0.023)		

Note: "Total effect (reduced form)" estimates, columns 1 and 5: Each cell reports a local average treatment effect from a separate regression, estimated using multi-site FRD methods. The dependent variable is described in the row label. The specification includes four endogenous treatment variables: (i) an indicator = 1 if the student took a second math class in 6th grade (the reported coefficient), (ii) an indicator = 1 if the student's regular math class in 6th grade was an advanced or honors section, (iii) the jackknife mean of baseline math test score among the student's regular math class peers in 6th grade, and (iv) the jackknife standard deviation of baseline math score among the same peers. The excluded instruments are a vector of indicator variables formed by the interaction of (i) the standard indicator variable = 1 if the student scored below the cut-score on the 5th grade math test, and (ii) school-specific indicators. Additional included regressors include: (i) a linear term for the forcing variable (5th grade test score), (ii) an interaction of the indicator for scoring below the cut-score and forcing variable, and (iii) school fixed effects. Estimation is by limited information maximum likelihood using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. The estimation sample is restricted to students who were observed in 6th, 7th, and 8th grades. Standard errors are clustered at the discrete values of the forcing variable.

"Marginal effect (indirect FRD)" estimates, columns 2 and 6: Each cell reports the LATE estimates using the "indirect FRD" method described in detail in the text which explicitly models the dynamic nature of treatment assignment up through 8th grade. Each of the five terms in equation 6 is estimated using the multi-site FRD methods described for columns 1 and 5, and then plugged into equation 6. Standard errors are calculated using the delta method. To obtain the complete covariance matrix, the constituent terms are estimated simultaneously using the "stacked equation" approach described in the text.

Columns 3 and 7 report outcome variable means (standard deviations) for the estimation sample.

* indicates $p<0.05$, ** $p<0.01$

Table 9—Effect of second, remedial math class on non-math outcomes

	5th grade math test score below …							
	level 1/2 cut-score				level 2/3 cut-score			
	Total effect (reduced form)	Marginal effect (indirect FRD)	Estimation sample mean (sd)	Student obs.	Total effect (reduced form)	Marginal effect (indirect FRD)	Estimation sample mean (sd)	Student obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reading score in grade 6	-0.027		-0.343	18828	0.003		0.085	35591
	(0.056)		(0.690)		(0.034)		(0.644)	
Successful transition from 9th to 10th grade	-0.008	-0.008	0.904	17898	0.017	0.015	0.946	33818
	(0.029)	(0.029)			(0.015)	(0.015)		
Two years foreign lang. by the end of high school	0.068	0.064	0.623	8643	-0.102*	-0.105*	0.729	17285
	(0.056)	(0.056)			(0.048)	(0.049)		
Number of years (0-4) during high school with class in								
Physical education	-0.012	-0.020	1.527	8643	0.024	0.037	1.546	17285
	(0.102)	(0.102)			(0.100)	(0.100)		
Music or arts	-0.113	-0.120	1.399	8643	0.006	0.019	1.579	17285
	(0.155)	(0.156)			(0.118)	(0.118)		
Graduated from high school in four years	0.058	0.059	0.804	5697	0.047	0.047	0.887	11226
	(0.053)	(0.053)			(0.050)	(0.050)		

Note: "Total effect (reduced form)" estimates, columns 1 and 5: Each cell reports a local average treatment effect from a separate regression, estimated using multi-site FRD methods. The dependent variable is described in the row label. The specification includes four endogenous treatment variables: (i) an indicator = 1 if the student took a second math class in 6th grade (the reported coefficient), (ii) an indicator = 1 if the student's regular math class in 6th grade was an advanced or honors section, (iii) the jackknife mean of baseline math test score among the student's regular math class peers in 6th grade, and (iv) the jackknife standard deviation of baseline math score among the same peers. The excluded instruments are a vector of indicator variables formed by the interaction of (i) the standard indicator variable = 1 if the student scored below the cut-score on the 5th grade math test, and (ii) school-specific indicators. Additional included regressors include: (i) a linear term for the forcing variable (5th grade test score), (ii) an interaction of the indicator for scoring below the cut-score and forcing variable, and (iii) school fixed effects. Estimation is by limited information maximum likelihood using a rectangular kernel and bandwidth of 23 scale score points above/below the cut-score. The estimation sample is restricted to students who were observed in 6th, 7th, and 8th grades. Standard errors are clustered at the discrete values of the forcing variable.

"Marginal effect (indirect FRD)" estimates, columns 2 and 6: Each cell reports the LATE estimates using the "indirect FRD" method described in detail in the text which explicitly models the dynamic nature of treatment assignment up through 8th grade. Each of the five terms in equation 6 is estimated using the multi-site FRD methods described for columns 1 and 5, and then plugged into equation 6. Standard errors are calculated using the delta method. To obtain the complete covariance matrix, the constituent terms are estimated simultaneously using the "stacked equation" approach described in the text.

Columns 3 and 7 report outcome variable means (standard deviations) for the estimation sample.

* indicates $p<0.05$, ** $p<0.01$