THE MEDIUM-RUN EFFECTS OF FLORIDA'S TEST-BASED PROMOTION POLICY

Marcus A. Winters
(corresponding author)
Department of Leadership, Research and Foundations
University of Colorado–Colorado Springs
Colorado Springs, CO 80918
mwinters@uccs.edu

Jay P. Greene
Department of Education Reform
University of Arkansas
Fayetteville, AR 72701
jpg@uark.edu

Abstract

We use a regression discontinuity strategy to produce causal estimates for the effect of remediation under Florida's test-based promotion policy on multiple outcomes for up to five years after the intervention. Students subjected to the policy were retained in the third grade, were required to be assigned to a high-quality teacher during the retained year, and were required to attend summer school. Exposure to these interventions has a statistically significant and substantial positive effect on student achievement in math, reading, and science in the years immediately following the treatment. But the effect of the treatment dissipates over time. Nonetheless, we find that the effect of remediation under the policy on academic achievement is statistically significant and of a meaningful magnitude several years after the student is exposed to the intervention. Though we cannot completely separate the differential effects of the treatments attached to the policy, we provide some evidence that assignment to a higher-quality teacher in the retained year is not the primary driver of the policy's effect.

1. INTRODUCTION

Several states and school districts currently employ remediation policies requiring low-performing students in particular grades to attend summer school and be retained in the grade if they fail to demonstrate possession of a predetermined skill level on one or more standardized assessments. Such policies are primarily intended to curtail "social promotion," the longstanding practice of promoting students to the next grade level for purposes of socialization even if they have not demonstrated an adequate level of proficiency for academic promotion. Most notably, such test-based promotion policies are operating in the Florida, Texas, New York City, and Chicago public school systems. In addition, Oklahoma, Arizona, and Indiana adopted similar programs in the last legislative session, and other states are reported to be seriously considering such policies.

Test-based promotion policies are particularly controversial because they can substantially increase the percentage of students who are retained in grade: the percentage of third graders retained in Florida increased to about 12 percent in the first two years of the state's adoption of the policy, up from only about 3 percent in the two years before the policy was adopted (Greene and Winters 2009). School systems adopting policies that dramatically increase grade retention do so despite a large body of research finding that retention is harmful to student achievement (for literature reviews expressing this view of the research, see Holmes 1989 and Jimerson 2001).

However, the believability of prior research showing that retention harms student achievement has been called into question (see, e.g., Greene and Winters 2007; Allen et al. 2009; and Hughes et al. 2010). Of the twenty-two articles evaluating the effect of grade retention on achievement from 1990 through 2006 that were identified in a recent meta-analysis by Allen et al. (2009), only six could be defined as high quality, meaning they included comparison groups with similar observed characteristics at baseline and adequate statistical controls. The meta-analysis discovered that higher-quality studies report more positive effects from grade retention than do lower-quality evaluations. Nonetheless, even these higher-quality articles do not tend to find that retention leads to substantial academic improvements.

Notable among this prior research is a series of studies that utilizes a regression discontinuity identification strategy (Greene and Winters 2007; Jacob and Lefgren 2004, 2007; Roderick and Nagaoka 2005). Such articles deserve particular attention because, unlike even very sophisticated matching strategies also characterized as high-quality designs by Allen et al. (2009), under minimal assumptions regression discontinuity accounts for both observed and unobserved characteristics related to both the likelihood that a student is

retained and his or her later academic achievement (van der Klaauw 2002; Imbens and Lemieux 2008).

In addition to concerns about its quality, much of the prior research evaluating the impact of retention and other related remediation policies is limited by the short-run nature of its findings. Prior studies tend to follow retained students for only one or two years after the intervention. Only two of the six studies characterized by Allen et al. (2009) as employing a high-quality design evaluated student performance more than two years after retention (Rust and Wallace 1993; Jimerson et al. 1997). Further, both of those studies utilized a matching design, which relies entirely on a set of observed characteristics and thus may not account for unobserved student factors related to retention. Though too recent to have been included in the meta-analysis, Hughes et al. (2010) use a propensity score-matching technique in order to look at the effect of retention in the first grade on student achievement four years later. The findings from these recent evaluations using matching techniques for identification suggest that retention has a positive effect in the short run that fades over time, often to the point of statistical insignificance.

The current article expands on previous research evaluating the effect of Florida's test-based promotion policy by considering the sustained effect of its treatment (Greene and Winters 2007). We utilize a regression discontinuity design to provide causal estimates of the relationship between a student having been remediated under the state's policy and achievement through the seventh grade. We map the academic performance of multiple student cohorts on both high- and low-stakes reading and math exams. We also provide estimates of the effect of a student being remediated under the policy on her achievement on an elementary science exam.

We find that students remediated under Florida's test-based promotion policy made substantial academic gains in all subjects in the short term and that these academic gains declined as the students progressed through school. However, we find a statistically significant and meaningful difference in student achievement several years after treatment.

Like other recently adopted policies, Florida's test-based promotion policy includes treatments other than retention. Retained students also attended summer school prior to the retained year and were required to be assigned to a high-quality teacher during the retained year. Unfortunately we cannot completely disentangle the effects of particular treatments under the policy, so our results are best thought of as the average effect of the entire remediation treatment. However, we do provide some evidence that the estimated effect of treatment does not appear to be driven by teacher assignment during the retained year.

We refer to our results as medium-run effects because Florida's policy has not been in existence long enough to allow for an estimate of its effect on high school graduation or later life outcomes. This is an important limitation because even if remediation leads to sustained academic improvements, we might expect retained students to be less likely than their socially promoted peers to graduate because they are older during their high school grades (Jacob and Lefgren 2007). Further, a recent study by Babcock and Bedard (2011) finds evidence that an increase in early grade retention within a state is related to increases in mean male hourly wages.

Nonetheless, a consideration of the policy's medium-run effects is of substantial policy interest. With many thousands of students currently subjected to such policies and other states contemplating similar programs, a consideration of the impact of Florida's program thus far provides important information for the policy discussion.

2. DATA

We utilize a rich data set made available by the Florida Department of Education's K-20 Data Warehouse. The data set contains test score and demographic information for the universe of test-taking students in Florida public schools in grades 3–8 from 2002–3 through 2008–9. The data set also includes a unique student identifier that allows us to track individual student performance over time.

We impose no restrictions on the data in addition to those used to develop the sample for the regression discontinuity analysis. The sample restrictions and descriptive statistics for the treatment and comparison groups are provided in a later section.

Florida's Test-Based Promotion Policy

Florida's test-based promotion policy was among a series of reforms adopted under the governorship of Jeb Bush. Students who entered the third grade in the fall of 2002 were the first subjected to the mandate. The law has applied to all subsequent cohorts of third-grade students in the state.

Florida's policy requires third-grade students to score at or above level 2 (the second lowest of five levels) on the state's high-stakes reading exam, the Florida Comprehensive Assessment Test (FCAT), in order to be default promoted to the next grade level. The benchmark score necessary to reach level 2 has remained consistent over time. Performing below the test score threshold can be said to influence default promotion only because students can receive one of several exemptions and be promoted despite their low performance. In fact, nearly half the students with test scores below the threshold in the policy's first year were promoted (Greene and Winters 2009).

Table 1. Description of Cohorts

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
2002–3	T1 C1					
2003–4	T1 T2 C2	C1				
2004–5	T2 T3 C3	T1 C2	C1			
2005–6	T3 T4 C4	T2 C3	T1 C2	C1		
2006–7	T4 T5 C5	T3 C4	T2 C3	T1 C2	C1	
2007–8	T5	T4 C5	T3 C4	T2 C3	T1 C2	C1
2008–9		T5	T4 C5	T3 C4	T2 C3	T1 C2

Notes: C = Control group, not retained; T = treatment group, retained. Numbers indicate cohort.

Students retained according to the policy are also subjected to additional interventions during the retained year. They are required to attend a summer reading camp prior to the retained year. In addition, during their retained year the schools are required to assign these students to a high-quality teacher as determined by performance data and above-satisfactory performance appraisals. Schools must also develop academic improvement plans that address the specific needs of these students during the retention year.

Because all students who were retained due to the policy also received the additional treatments, we are not able to distinguish the effect of retention itself from that of these other interventions. Interpreting our results as the effect of retention is particularly difficult given the large body of research showing that teacher quality has a substantial effect on student achievement (see, e.g., Rivkin, Hanushek, and Kain 2005) as well as prior work showing the benefits from summer school (Jacob and Lefgren 2004).

Because the analysis cannot disentangle the effects of retention from the effects of these other interventions, our primary estimates should be considered an average treatment effect of remediation under Florida's test-based promotion policy.

Student Cohorts and the Special Case of Cohort 1

The fact that Florida's policy has been in effect for seven years allows us to evaluate the policy's influence for multiple cohorts and over a sustained period of time. We estimate the effect of remediation under Florida's test-based promotion policy on these cohorts both individually and as a group.

Table 1 tracks the movement of the cohorts under consideration through grade levels over time. T indicates the group of students who were retained (treated), and C indicates the group of students who were promoted at the end of the third grade (control). The number next to the letter indicates the

student's cohort. For instance, cohort 1 first entered the third grade in 2002–3, and cohort 2 first entered the third grade in 2003–4.

As the table shows, the farthest our data set allows us to follow a cohort (cohort 1) is eighth grade, because it is the last grade in which both retained and promoted groups are observed. The youngest cohort evaluated is cohort 5 (fourth-grade achievement), which is the last our data set allows us to observe for both the promoted and the retained group.

Cohort 1, the first subjected to the policy, represents a special case. As table 1 shows, all cohorts subsequent to cohort 1 share a grade level with a group of students in another cohort. For instance, when retained students in cohort 3 (T3) are in fourth grade they share classrooms with promoted students from cohort 4 (C4). Because cohort 1 was the first to be subjected to the policy, students from that group who were promoted to fourth grade do not share classrooms with a large group of students who were retained in third grade from a previous cohort. An implication of this phenomenon is that the quality of peers sharing classrooms with promoted students from cohort 1 in subsequent years is different than for students who were retained from that group.

Essentially, students who were promoted to fourth grade at the end of 2002–3 no longer shared classrooms with a large number of very low-performing students who were instead retained in third grade. About 14 percent of this third-grade class was retained in third grade. Subsequent cohorts subjected to the policy were also removed from many of their low-performing classmates, but these low-performing students were replaced in the cohort by students who were retained at the end of the previous year.

Promoted students in cohort 1 (C1) attend later grades with higher-average-quality peers than do students in subsequent cohorts. Prior research suggests that peer quality influences student learning gains during the school year (see, e.g., Hanushek et al. 2003). Since the promoted group from cohort 1 has higher-quality peers on average in later grades than does the retained group from cohort 1, the estimated influence of retention on test scores at the end of later grades are likely biased downward.

Though they cannot be generalized to the overall effect of treatment under test-based promotion policies, estimates of the treatment effect within cohort 1 are relevant for policy. The experience of the first cohort subject to a test-based promotion policy that is adopted by another school system would have an experience similar to our cohort 1. However, all subsequent cohorts in Florida (and anywhere a similar policy is adopted) might be expected to have a different experience than the first cohort subject to the policy.

We address the special case of cohort 1 in two ways. First, when estimating models that incorporate students from multiple cohorts, we report only the

results of models that exclude students from cohort 1.[1] We also estimate the effect of retention on each individual cohort over time and keep in mind the differing peer effects when comparing estimates resulting from cohort 1 with those from other cohorts.[2]

3. IDENTIFICATION STRATEGY

We employ a regression discontinuity identification strategy in order to provide causal estimates of the effect of remediation under Florida's test-based promotion policy on later student achievement. Regression discontinuity is applicable in cases where assignment of a treatment is either entirely or partially a function of whether an individual falls above or below a continuously measured benchmark. When certain assumptions are satisfied, the procedure closely approximates a randomized experiment (van der Klaauw 2002; Imbens and Lemieux 2008).

We take advantage of Florida's use of a discrete benchmark on the third-grade reading test to help determine whether a child is subjected to remediation. Though their reading achievement is essentially equivalent, students with reading scores that fell just below the eligibility threshold were far more likely to be remediated than were students whose scores were just above the eligibility threshold.

Our analyses include only observations for students whose test score in their initial third-grade year fell within a small neighborhood around the passing cutoff—a score of 1046 on the test's developmental scale. We define the neighborhood as scores within 18 points below the cutoff or 23 points above the cutoff. We choose these points on the distribution because it is recommended that the analysis include no fewer than four points on the distribution on either side of the cutoff; since the scale on the third-grade assessment only allows for certain values, these are the four nearest observation points on either side of the cutoff (Schochet et al. 2010). Use of this narrowest band possible improves the likelihood that the above and below groups are equivalent on both observed and unobserved characteristics, and our statewide data set allows for a large enough number of student observations to produce precise estimates within even this very narrow bandwidth.

Table 2 reports the number and percentage of students earning each observed initial third-grade test score who were retained or promoted the

1. Results are generally similar in models that include cohort 1.
2. Notice that exclusion of cohort 1 students during estimation does not remove their influence on peer quality for students in cohort 2, because the students still did in fact share classrooms and thus played a role in the education production process determining cohort 2's achievement that is under consideration.

Table 2. Number and Percent Retained for Students with Particular Third-Grade Reading Scores in Sample

	Initial Third-Grade Reading Score	*Below Threshold* 1027	1033	1039	1045	*Above Threshold* 1051	1057	1063	1069	All Third Grade
	Number promoted	433	443	482	504	851	859	920	982	151,106
Cohort 1	Number retained	346	300	341	333	40	54	33	37	24,334
	Percent retained	44%	40%	41%	40%	4%	6%	3%	4%	14%
	Number promoted	508	535	542	603	805	833	829	847	173,836
Cohort 2	Number retained	220	190	219	168	21	38	30	22	18,664
	Percent retained	30%	26%	29%	22%	3%	4%	3%	3%	10%
	Number promoted	472	528	540	578	836	779	862	843	172,842
Cohort 3	Number retained	253	242	216	207	30	33	31	27	15,942
	Percent retained	30%	26%	29%	22%	3%	4%	3%	3%	8%
	Number promoted	379	364	422	389	650	695	664	675	178,655
Cohort 4	Number retained	173	176	187	183	22	28	26	25	11,421
	Percent retained	31%	33%	31%	32%	3%	4%	4%	4%	6%
	Number promoted	495	550	553	586	700	733	793	838	169,163
Cohort 5	Number retained	146	151	135	137	14	31	24	19	11,604
	Percent retained	23%	22%	20%	19%	2%	4%	3%	2%	6%

following year. The table makes apparent that for each cohort a student with a score below the threshold for default promotion (1046 on the FCAT developmental scale) was far more likely to be retained than was a student whose score fell just above this cutoff.

The final column of table 2 also shows all third-grade students for whom we have test score data who were retained or promoted in particular years. It is clear that the percentage of third graders being retained has decreased substantially over the life of the policy. Much of that decline has to do with a general increase in student performance on the reading test across the state. However, such general academic improvements in the state do not hinder our estimation, particularly those that are focused on only a single cohort.[3] Though the distribution of third-grade student test scores has shifted, there are still many students with academic achievement very near the threshold for promotion. Further, the table also shows that the percentage of students who are retained from our restricted group of students just above and below the threshold has remained relatively consistent, particularly for cohorts 2, 3, and 4.

It is worth noting that the high internal validity of the regression discontinuity procedure potentially comes at the price of external validity. Our analysis tells us a great deal about the effect of remediation on the later achievement of students whose third-grade score was within a narrow band of the eligibility cutoff. However, it is possible that the treatment effect could differ—either positively or negatively—for students whose score was farther from the eligibility cutoff.

That students and schools know about the policy and the exact cutoff score for default promotion is worrisome, because the regression discontinuity procedure will not produce causal estimates if the subjects manipulate whether they fall just above or below the threshold (Schochet et al. 2010). However, though we suspect that both students and schools likely respond to the policy by attempting to push past the threshold for promotion, we argue that this is not a particular problem for our estimation. There is no reason to believe a student's standardized reading score is not a valid and reliable measure of the student's true proficiency level. Further, as seen by the table, we do not observe a heavy clustering of students just above the threshold for promotion, which would be expected where the forcing variable has been manipulated.

3. The general improvements in Florida test scores over time could potentially affect the interpretation of the estimates from models that include multiple cohorts. In particular, the change in test scores makes the consideration of local regressions around the eligibility thresholds somewhat unclear. Further, the general test score improvements imply that later cohorts are surrounded by higher-quality peers as they progress through school, which could influence the estimates. Nonetheless, that the results appear to be stable when we consider individual cohorts suggests that the rightward drift in test scores over time is not substantially influencing our estimates.

Table 2 also shows that there are many students with scores below the threshold who were promoted and some students with scores above the threshold who were nonetheless remediated. Because exposure to the interventions is not strictly determined by where the student's score falls relative to the threshold for default promotion, we utilize the "fuzzy" regression discontinuity strategy. Essentially this strategy uses an indicator for whether the student's score is below the threshold as an instrumental variable for remediation in a two-stage least squares approach.

There are two stages in the estimation. The first utilizes observed characteristics about the student, including her test score and an indicator for whether it falls above or below the eligibility threshold, in order to predict the likelihood that she is remediated. Only observations of students in the third grade whose score falls within the previously defined neighborhood of the cutoff score are utilized in this first stage regression. Formally:

$$remediated_{is3} = \beta_0 + \beta_1 X_{is3} + \beta_2 Below_{is3} + \gamma_s + \varepsilon_{is3}, \quad (1)$$

where $remediated_{is3}$ equals one if student i enrolled in school s was remediated at the end of his third-grade year;[4] X is a series of observed characteristics about the student, including race/ethnicity, gender, an indicator for eligibility for free or reduced price lunch, an indicator for being identified as having a disability, and scores on the third-grade high- and low-stakes reading and math tests (that is, four tests in total), including the FCAT reading score that is linked to the retention policy. Below the instrumental variable is an indicator that equals one if the student's FCAT reading score in the third grade was below the eligibility threshold set by the policy and zero if the score was above the threshold; γ is a fixed effect for the student's school; ε is a stochastic term clustered by school; and β_0 through β_2 are parameters to be estimated.

Equation 1 is estimated via ordinary least squares (OLS), which results in a linear probability model (LPM). Though "remediated" is a dichotomous dependent variable, we utilize the LPM in this case rather than a probit model because inclusion of more than a thousand school fixed effects makes conversion difficult with probit, which is estimated via maximum likelihood. We estimate equation 1 independently for students in each third-grade cohort from 2002–3 through 2006–7.

The coefficients from equation 1 are used to estimate the probability that the child received the remediation treatments under the test-based promotion policy conditional on observed characteristics, which we write as $\hat{r}$ and capture

4. A student is classified as having been remediated if he is again observed in the third grade in the next year.

for each student. We then utilize $\hat{r}$ in a second-stage OLS regression evaluating the student's test score in a subsequent year. Formally:

$$Y_{isg} = \alpha_0 + \alpha_1 X_{isg} + \alpha_2 Y_{is3} + \alpha_3 \hat{r}_i + \delta_s + \mu_{isg}, \quad (2)$$

where Y is the student's test score; δ represents a school fixed effect; μ is a stochastic term clustered by school; and α_0 through α_3 are parameters to be estimated. If we believe the identification assumptions (tested in the next section), the estimate of α_3 can be interpreted as the causal influence of remediation under the test-based promotion policy on student academic achievement. Notice that since the probability that the child received treatment does not change over time, $\hat{r}$ has the same value for the student regardless of the grade-level test score under consideration.

The index g represents the student's grade in the year under consideration. Note that the student's test score is used as a regressor on the right-hand side of equation 2 always represents the student's test score in the initial third-grade year.[5] We specify the student's prior achievement in this way for two reasons. First, the regression discontinuity procedure requires that the regression model control for the point system used to make the student subject to treatment, and this "forcing variable" in our case is the student's score on the third-grade FCAT reading exam (Schochet et al. 2010). Second, we are interested in evaluating the effect of the treatment applied in the retained year on a student's achievement in grades several years later. If we estimated a common value-added model in which we accounted for the student's test score in the prior year, the results would be difficult to interpret because the student's previous test score would also be partly determined by the treatment. Thus our procedure is to identify treatment and control groups that had very similar test scores at the end of their shared third-grade year and then evaluate whether and to what extent their test scores in subsequent grades differ relative to one another.

Equation 2 is estimated for students at the same grade level, not the same year. The performance of retained students is compared with that of other students with whom they attended the third grade for the first time. Since students in the treatment group were retained in the third grade, most of them are a grade level behind their original third-grade classmates in each subsequent year. There is something of a debate in the prior research on grade retention over whether comparisons between retained and promoted students should be made within grade or within year. We follow

5. Though retained students are in the third grade twice, we utilize as the control in equation 2 their initial score in that grade, which contributed to their retention.

the within-grade approach and compare the performance of retained and promoted students when they attended the same grade level, though for the treated group the attendance in that grade comes after an additional year of instruction.

Like some other prior researchers, we argue that the within-grade comparison is the most policy relevant because it most aligns with what schools are interested in: the student's performance relative to his same-grade peers (see, for instance, Alexander, Entwisle, and Dauber 1994). In addition, we point out that if we think of schooling in the long-term context, the additional year of schooling itself is one of the most important interventions under consideration. The ultimate question facing those interested in policies that include a retention component is whether retained students acquire a greater level of proficiency by the time they leave the school system than they would have without the intervention in the elementary grade. One potential benefit from retention is the additional year of schooling. If both groups graduate after four years in high school, a retained student who completes high school will have received one more year of schooling than a student who was not retained. The ultimate relevant comparison between those students is their final test score in the twelfth grade, not their test score nine years after the third grade. Consequently, at each point in their academic careers it is the within-grade comparison that is of most interest.

We estimate various forms of equation 2. Each model is restricted to students observed in the particular grade level under consideration. When estimating models that combine cohorts into a single regression, we include only cohorts for which we observe both the treated and the control group in a particular grade. For example, recalling table 1, the combined estimation of the effect of treatment on student performance in grade 6 can include only students from cohorts 2 and 3 (recall that cohort 1 is excluded in all models that combine cohorts), while the combined estimation of the effect of treatment on student performance in grade 4 includes all five cohorts. We also report the results of models restricted to individual cohorts in order to follow their progress as they enter later grades.

We estimate models that alter the dependent variable to include the student's score on one of five standardized tests. We analyze math and reading scores from two different assessments: the FCAT, which is used for different facets of the state's accountability system, and the NRT (the norm-referenced test, a low-stakes test that is a version of the Stanford 10 and was administered to all Florida public school students in grades 3–10 through the 2007–8 school year before its use was discontinued. Finally, we also evaluate student achievement on the state's fifth-grade science exam, which is the only grade in which the exam is given that we observe for both retained and promoted

students.[6] In order to ease interpretation, we standardize scores on all exams by grade and year to have a mean of zero and standard deviation of 1.

4. TESTING THE IDENTIFICATION ASSUMPTIONS

The first assumptions to consider are those linked to any instrumental variable approach. In particular, the classical assumptions necessary to use Below Threshold as an instrument are that it be correlated with receipt of treatment but otherwise unrelated to later student achievement.

The descriptive statistics provided in table 2 show that Below Threshold meets the first condition of being highly correlated with receipt of the treatment. It is clear from the table that those with scores below the threshold are far more likely to be retained than those with scores just above it. Further, though not reported for space considerations, when estimating equation 1 we find that for each cohort there is a statistically significant relationship between whether the student's score falls below the threshold for default promotion and the likelihood that the student is retained.[7]

The regression discontinuity procedure allows us to assume that Below Threshold meets the second condition of an instrumental variable: that it be unrelated to later student outcomes except through its influence on the probability that a student receives treatment. The idea behind the regression discontinuity procedure is that for students with scores very close to the eligibility threshold, randomness plays a meaningful role in determining whether the student's score falls just above or just below the cutoff. If that is the case, then whether the score of a student in the sample falls below the cutoff is unrelated to the student's later test score except through its influence on the probability that the student received treatment.

The use of variables like Below Threshold as instruments in a regression discontinuity framework is widely accepted in the literature. We can also test the validity of this assumption by evaluating whether the observed baseline characteristics of students in our sample whose scores were above and below the threshold are statistically identical.

Consistent with the identifying assumption, the results of this comparison for all cohorts, shown in table 3, show few significant differences in the

6. The test is also administered in the eighth grade. Though we observe both promoted and retained students from cohort 1 in the eighth grade, we do not include this analysis because of the special nature of cohort 1, as described in an earlier section.
7. We also estimated models that changed the functional form of the forcing variable by adding polynomials of it as regressors, but this produced no meaningful difference in any results. We also tested models that incorporated an interaction between the forcing variable and Below Threshold, but this did not produce a meaningful change either. We thus utilize the more parsimonious model in our analysis, which treats the relationship between the forcing variable and retention as linear.

Table 3. Demographics of Students Above and Below Test Score Threshold

		White	African American	Hispanic	Lunch: Applied, Not Eligible	Eligible for Free Lunch	Eligible for Reduced Price Lunch	IEP
Cohort 1	Above threshold	36.6%	36.0%	26.0%	3.1%	55.1%	10.7%	31.2%
	Below threshold	34.0%	35.6%	26.6%	2.5%	56.0%	11.0%	32.6%
Cohort 2	Above threshold	33.1%*	37.6%	25.1%*	2.6%	59.0%	10.1%	34.5%*
	Below threshold	30.2%*	36.6%	29.0%*	2.7%	61.3%	10.6%	37.2%*
Cohort 3	Above threshold	37.2%	33.5%	30.3%	3.3%	60.1%*	10.5%*	36.3%
	Below threshold	31.4%	35.4%	29.0%	2.7%	62.5%*	8.7%*	37.8%
Cohort 4	Above threshold	27.5%	38.7%	28.6%	2.6%	63.5%	10.1%	39.2%*
	Below threshold	28.4%	38.7%	28.7%	2.2%	64.5%	10.0%	42.3%*
Cohort 5	Above threshold	28.1%	35.7%	31.0%	3.0%	61.0%	11.7%	32.1%*
	Below threshold	27.4%	37.5%	30.8%	3.0%	63.2%	10.9%	35.4%*

Note: IEP = individualized education program
*$p < 0.05$

observed characteristics of the two groups. The main exception is that students with scores just below the threshold appear to be slightly more likely to have a disability than students just above the threshold. The few other exceptions where the groups are statistically different are not meaningful in size and tend to disadvantage the students with scores below the threshold.

These results show that those who fell just above and below the eligibility threshold were essentially identical in their observed characteristics. This finding allows us to also assume that the groups are identical in their unobserved characteristics as well. Our use of the narrowest threshold possible to determine our estimation samples strengthens the case that Below Threshold meets this exogeneity condition.

Since we are concerned with the effects of the treatment over time, an additional factor to consider regarding the validity of the comparison has to do with attrition from the sample, which would be alarming if it were disproportionate among the treatment and comparison groups. Students could leave our sample if they moved to a school outside the Florida public school system, dropped out, or for some reason did not have an observable test score in a later year.

Table 4 maps the attrition for the treatment and control groups for each of the five cohorts. Notice that for each cohort we observe 100 percent of the remediated students the year after the initial third-grade year. That is because we can identify whether a student has been retained only if he is observed in the subsequent year. At each point it appears that attrition is relatively similar for both the retained and promoted students for each cohort. If we were to characterize a difference, it appears that there was slightly less attrition for remediated students.

5. RESULTS

We first consider the results of models that combine our cohorts into a single estimation. The tables make clear that because we do not observe all cohorts in each grade level, not all cohorts are utilized in each estimation. Since at this point we are focused on models that combine cohorts, we report only the results of models that utilize at least two cohorts for estimation. Finally, as discussed previously, because students in cohort 1 represent a special case they are excluded from all models that combine cohorts.

Table 5 reports the results from our estimate of the impact by grade on reading test scores of remediation under the test-based promotion policy. The first set of results reports estimates when a school fixed effect is not included, and the second set adds a school fixed effect into the estimation. For these models that utilize multiple cohorts, we prefer the estimates that incorporate a school fixed effect, though results are similar across the specifications.

Table 4. Attrition by Cohort and Promotion Status

Year	2002–3	2003–4	2004–5	2005–6	2006–7	2007–8	2008–9
Cohort 1							
Promoted	5,669	5,150	5,052	4,829	4,665	4,595	4,415
Percent of beginning		91%	89%	85%	82%	81%	78%
Retained	1,558	1,558	1,485	1,437	1,368	1,317	1,282
Percent of beginning		100%	95%	92%	88%	85%	82%
Cohort 2							
Promoted		5,621	5,233	5,100	4,883	4,761	4,648
Percent of beginning			93%	91%	87%	85%	83%
Retained		974	974	921	889	849	817
Percent of beginning			100%	95%	91%	87%	84%
Cohort 3							
Promoted			5,642	5,206	5,034	4,834	4,689
Percent of beginning				92%	89%	86%	83%
Retained			1,096	1,096	1,044	1,010	971
Percent of beginning				100%	95%	92%	89%
Cohort 4							
Promoted				4,436	4,041	3,910	3,780
Percent of beginning					91%	88%	85%
Retained				879	879	830	802
Percent of beginning					100%	94%	91%
Cohort 5							
Promoted					5,507	5,105	4,976
Percent of beginning						93%	90%
Retained					718	718	670
Percent of beginning						100%	93%

On both the FCAT and the NRT exams we find evidence that students make substantial improvements in reading proficiency in the grades immediately following remediation. When remediated students are in the fourth grade they perform about a third of a standard deviation better than did socially promoted students when they were in the fourth grade. It is also clear from the regressions that the influence of remediation on student reading scores tends to decline with each subsequent grade. Nonetheless, by the sixth grade—four years after the retained student's initial third-grade year—remediation is related to about a 0.168 standard deviation increase in reading scores.

Our findings from math are similar to those reported for reading. Table 6 shows that remediation under the test-based promotion policy is related to an early substantial increase in math test scores and tends to decline each year. However, we again see a statistically significant and relatively substantial influence from remediation of about a fifth of a standard deviation when

Table 5. Regression Results by Grade: Reading

				FCAT		
	Grade 4	**Grade 5**	**Grade 6**	**Grade 4**	**Grade 5**	**Grade 6**
Retention (predicted)	0.340*	0.236*	0.223*	0.292*	0.168*	0.190*
	[0.0218]	[0.0286]	[0.0335]	[0.0251]	[0.0341]	[0.0371]
				NRT		
Retention (predicted)	0.268*	0.231*		0.190*	0.131*	
	[0.0262]	[0.0327]		[0.0312]	[0.0437]	
Student controls	√	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√	√
Treated cohort 1						
Treated cohort 2	√	√	√	√	√	√
Treated cohort 3	√	√	√	√	√	√
Treated cohort 4	√	√		√	√	
Treated cohort 5	√			√		
School fixed effect				√	√	√

Notes: Dependent variable is student's score on the FCAT or NRT reading test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.01$

students are in the sixth grade. Results are similar on both the high-stakes FCAT exam and the low-stakes NRT.

Table 7 reports the results of regressions looking at whether the relationship between remediation and achievement in math or reading differed by student gender or race/ethnicity. In both subjects we see little evidence that the effect of treatment differs meaningfully according to such demographic characteristics.

We now consider how remediation influenced the achievement of students from each cohort separately. We do not incorporate a school fixed effect in these models because in most cases the more limited sample did not allow for enough observations of students within particular schools to be estimated confidently.[8]

Table 8 reports the results of regression models evaluating the effect of the treatment on reading achievement on the FCAT for each cohort by grade. The results indicate that remediation under the test-based promotion policy has an early large effect on student reading test scores that tends to fade over time. However, for each cohort except cohort 1, the effect remains statistically significant in later years and is often quite substantial. Remediated students

8. In most cases we observe only an average of four students per school when the sample is restricted by cohort and grade.

Table 6. Regression Results by Grade: Math

	Grade 4	Grade 5	Grade 6	FCAT Grade 4	FCAT Grade 5	FCAT Grade 6
Retention (predicted)	0.423*	0.277*	0.208*	0.382*	0.278*	0.179*
	[0.0229]	[0.0274]	[0.0366]	[0.0267]	[0.0328]	[0.0390]
				NRT		
Retention (predicted)	0.262*	0.222*		0.160*	0.161*	
	[0.0290]	[0.0306]		[0.0373]	[0.0386]	
Student controls	√	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√	√
Treated cohort 1						
Treated cohort 2	√	√	√	√	√	√
Treated cohort 3	√	√	√	√	√	√
Treated cohort 4	√	√		√	√	
Treated cohort 5	√			√		
School fixed effect				√	√	√

Notes: Dependent variable is student's score on the FCAT or NRT math test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.01$

from cohort 2 outperform their untreated peers on the FCAT reading test in the seventh grade by about 0.183 standard deviations.

The results are similar in math. Table 9 shows that remediated students from cohort 2 perform about 0.174 standard deviations better than their untreated peers in seventh grade. We again see that the effects from retention tend to decline over time but remain statistically significant in all cases, this time including cohort 1. Though not reported here because of space considerations, the results in both math and reading are similar on the NRT exams.

Finally, table 10 reports estimates for the effect of remediation on student achievement under the test-based promotion policy on the state's fifth-grade science exam. The results show that the positive effects of remediation, at least in fifth grade, are apparent in the student's performance in subjects other than the core subjects of math and reading. For each cohort we find that remediation under the test-based promotion policy increases student science achievement in the fifth grade by between a fifth and a quarter of a standard deviation.

6. THE EFFECT OF TEACHER ASSIGNMENTS DURING THE RETAINED YEAR

Because remediated students were subjected to multiple interventions—retention, assignment to a high-quality teacher, and summer school attendance—our primary analysis cannot completely disentangle the effect

Table 7. Regression Results by Grade and Student Demographics: Reading and Math

	Reading			Math		
	Grade 4	**Grade 5**	**Grade 6**	**Grade 4**	**Grade 5**	**Grade 6**
			Male			
Retention (predicted)	0.337*	0.269*	0.227*	0.384*	0.258*	0.211*
	[0.0312]	[0.0397]	[0.0478]	[0.0307]	[0.0363]	[0.0497]
			Female			
Retention (predicted)	0.339*	0.189*	0.213*	0.465*	0.292*	0.200*
	[0.0294]	[0.0375]	[0.0448]	[0.0326]	[0.0393]	[0.0482]
			African American			
Retention (predicted)	0.368*	0.238*	0.276*	0.440*	0.259*	0.186*
	[0.0367]	[0.0475]	[0.0542]	[0.0385]	[0.0477]	[0.0589]
			Hispanic			
Retention (predicted)	0.334*	0.279*	0.211*	0.450*	0.297*	0.254*
	[0.0445]	[0.0539]	[0.0664]	[0.0474]	[0.0487]	[0.0743]
			White			
Retention (predicted)	0.327*	0.212*	0.190*	0.407*	0.292*	0.177*
	[0.0379]	[0.0502]	[0.0568]	[0.0365]	[0.0464]	[0.0598]
Student controls	√	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√	√
Treated cohort 1						
Treated cohort 2	√	√	√	√	√	√
Treated cohort 3	√	√	√	√	√	√
Treated cohort 4	√	√		√	√	
Treated cohort 5	√			√		

Notes: Dependent variable is student's score on the FCAT reading or math test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender (excluded from regressions evaluating male or female students), race/ethnicity (excluded from regressions evaluating students of particular race/ethnicity), free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.01$

of any particular piece of the policy on later student outcomes. The inability to separate the effects of these individual treatments is unfortunate, since determining which pieces of the policy are most and least effective could guide the decisions of policy makers considering similar policies. Given the desire to understand as much as possible about what is driving the substantial and sustained average treatment effect described in the prior section, in what follows we provide a test of the extent to which assignment to a high-quality teacher in the student's retained year is likely to be driving the treatment effect.

As discussed previously, according to the policy, retained students must be assigned to a high-quality teacher during their retained year. Presumably after the retention year remediated students are assigned to teachers in a

Table 8. Regression Results by Grade and Cohort: FCAT Reading

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
		Cohort 1			
Retention (predicted)	0.324**	0.168**	0.0942*	0.0052	0.0207
	[0.0329]	[0.0348]	[0.0384]	[0.0422]	[0.0442]
		Cohort 2			
Retention (predicted)	0.311**	0.224**	0.231**	0.183**	
	[0.0404]	[0.0447]	[0.0459]	[0.0491]	
		Cohort 3			
Retention (predicted)	0.292**	0.238**	0.210**		
	[0.0399]	[0.0499]	[0.0470]		
		Cohort 4			
Retention (predicted)	0.385**	0.327**			
	[0.0447]	[0.0490]			
		Cohort 5			
Retention (predicted)	0.409**				
	[0.0490]				
Student controls	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√
School fixed effect					

Notes: Dependent variable is student's score on the FCAT reading test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.05$; **$p < 0.01$

manner similar to their classmates. Thus in order to test for the extent to which the teacher assignment treatment is driving the average effect from the test-based promotion policy, we use our rich data set to estimate a series of models similar to those reported above but incorporating a fixed effect for the student's teacher in her final third-grade year. Thus students who were promoted after their initial third-grade year are matched to their only teacher in that grade, and students who were retained in the third grade are matched to their teacher during their repeating third-grade year.

Our student-level data set includes variables that match students to their teachers in particular classrooms. However, many third-grade students are attached to more than one teacher during the school year. We develop a matching protocol in order to arrive at a single observation of a student in a single teacher's classroom during their final third-grade year for math and reading.[9]

9. First, we include only teachers listed as the head of a self-contained classroom. If students are still observed to be attached to multiple teachers, we assign them to particular course numbers. Students are first matched to the teacher in the course listed as third grade, and about 85 percent of students are matched to this teacher. Remaining students are matched to courses specific to elementary math or reading, depending on the analysis. For the reading sample, the progression

Table 9. Regression Results by Grade and Cohort: FCAT Math

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
		Cohort 1			
Retention (predicted)	0.413***	0.316***	0.150***	0.0872*	0.0888**
	[0.0373]	[0.0393]	[0.0432]	[0.0446]	[0.0450]
		Cohort 2			
Retention (predicted)	0.360***	0.268***	0.199***	0.174***	
	[0.0431]	[0.0470]	[0.0526]	[0.0525]	
		Cohort 3			
Retention (predicted)	0.368***	0.276***	0.217***		
	[0.0455]	[0.0461]	[0.0483]		
		Cohort 4			
Retention (predicted)	0.416***	0.328***			
	[0.0438]	[0.0506]			
		Cohort 5			
Retention (predicted)	0.523***				
	[0.0479]				
Student controls	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√
School fixed effect					

Notes: Dependent variable is student's score on the FCAT math test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Table 10. Regression Results by Cohort: FCAT Fifth-Grade Science Exam

	All Cohorts	All But Cohort 1	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Retention (predicted)	0.268*	0.216*	0.261*	0.229*	0.242*	0.256*
	[0.0222]	[0.0262]	[0.0416]	[0.0460]	[0.0511]	[0.0519]
Student controls	√	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√	√

Notes: Dependent variable is student's score on the FCAT fifth-grade science test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.01$

assigns the student (in order) to the teacher listed as language arts elementary, reading elementary, and finally language arts K–5. For the math sample, the progression next assigns the student to the teacher listed as math elementary, and then math K–5. About 96 percent of third-grade students in our data set are matched to a teacher according to these progressions, and remaining students are excluded from the analyses.

We first estimate a model identical to equation 1, except we add a squared term for the forcing variable, and we again capture the estimated probability that the student is remediated, $\hat{r}$. We then estimate a model identical to equation 2, except that it replaces the school fixed effect with a fixed effect for the student's final third-grade teacher, λ:

$$Y_{isg} = \theta_0 + \theta_1 X_{isg} + \theta_2 Y_{is3} + \theta_3 \hat{r}_i + \lambda_{i3} + \tau_{isg}. \quad (3)$$

A direct test for whether teacher assignments are driving the treatment effects reported previously would require that we estimate equation 3 using an identical sample to that used to estimate equation 2 previously. However, our use of such a narrow neighborhood around the eligibility threshold to identify the estimation sample limits the analysis to so few students across the state that only a very few students are observed with each third-grade teacher. When a third-grade teacher fixed effect is used on the primary sample, the average teacher in the analysis is matched to fewer than two students, making use of the teacher fixed effect problematic. Thus we cannot simply use this new model in an estimation using the previous sample. This is also why we do not include a teacher fixed effect in the prior estimation of the average treatment effect.

In order to develop a sample with enough teachers for realistic estimation of the treatment effect, we considerably expand the neighborhood around the eligibility cutoff used to identify the treatment and control groups. We report results from models that include students whose initial third-grade score was within one-tenth, one-quarter, and one-half of a standard deviation from the eligibility threshold. In comparison, students in the main analyses have initial third-grade test scores that are about 0.05 standard deviations from the eligibility threshold score. It is the expansion of the neighborhood around the threshold used to determine the sample that led us to include the squared term for the forcing variable in the first stage.

In order to avoid losing students and teachers to attrition, we only use these models to estimate the effect of remediation on fourth-grade achievement. We also only estimate models that combine student cohorts, because the number of teachers is again too small when the analysis is limited to a single cohort of students.

Though the sample differences make it impossible to directly compare the results from the estimation of equation 3 with that of equation 2, they still allow for a check on the extent to which teacher assignments are likely driving the previously reported average treatment effect of remediation. If the effect of the remediation policy is largely driven by assignment to a high-quality teacher in the retained year, we should expect that inclusion of the third-grade teacher fixed effect would substantially reduce the estimated average treatment effect

Table 11. Regressions Including Fixed Effect for Final Third-Grade Teacher

	Within Half Standard Deviation of Threshold					
	Math			***Reading***		
Retention (predicted)	0.435*	0.373*	0.349*	0.345*	0.287*	0.269*
	[0.0178]	[0.0166]	[0.0190]	[0.0157]	[0.0157]	[0.0181]
	Within Quarter Standard Deviation of Threshold					
Retention (predicted)	0.404*	0.347*	0.346*	0.312*	0.252*	0.237*
	[0.0185]	[0.0186]	[0.0253]	[0.0173]	[0.0184]	[0.0242]
	Within Tenth Standard Deviation of Threshold					
Retention (predicted)	0.455*	0.419*	0.427*	0.335*	0.292*	0.281*
	[0.0220]	[0.0243]	[0.0419]	[0.0203]	[0.0240]	[0.0405]
Student controls	√	√	√	√	√	√
Original third-grade test scores	√	√	√	√	√	√
Treated cohort 1						
Treated cohort 2	√	√	√	√	√	√
Treated cohort 3	√	√	√	√	√	√
Treated cohort 4	√	√	√	√	√	√
Treated cohort 5	√	√	√	√	√	√
School fixed effect		√			√	
Third-grade teacher fixed effect			√			√

Notes: Dependent variable is student's score on the FCAT reading or math test. Test scores are standardized by grade and year to have a mean of zero and standard deviation of 1. Student controls include gender, race/ethnicity, free or reduced price lunch eligibility status, and an indicator for whether student is disabled. Standard errors clustered by school are reported in brackets.
*$p < 0.01$

of remediation on the student's fourth-grade test scores. Thus in this check we report the results from estimation of both equations 2 and 3 using the expanded samples and are interested in whether the coefficient estimate on the predicted treatment is substantially different across these specifications.

The results from this analysis are reported in table 11. The results show that inclusion of the third-grade teacher fixed effect has almost no influence on the estimated effect of the policy treatment on student fourth-grade test scores. Thus it appears unlikely that the treatment effect identified in this article's main analysis is driven by the teacher to which a student was assigned during his retained year.

The results shown in table 11 also serve as a useful robustness check for our main findings. The first two columns for the math and reading analyses are equivalent to the models reported in tables 5 and 6, except that we have expanded the neighborhood surrounding the eligibility threshold used to derive the sample. The estimated effect of remediation on fourth-grade math and reading scores in models that incorporate either no fixed effect or only a school fixed effect are very similar to those reported in the previous models using the more restricted sample. Because they more plausibly account for unobserved

student heterogeneity, we continue to prefer the previously reported models as estimates of the average treatment effect of the remediation policy. However, that the estimates remain stable as the sample is expanded suggests that our results are not driven by the choice of where we draw the neighborhood around the eligibility threshold to choose our sample.

7. CONCLUSION

This article has evaluated the sustained effects of remediation on student achievement under Florida's test-based promotion policy. We find evidence that remediation under the policy has a large positive effect on student achievement in the years closely following treatment but that the magnitude of this effect declines over time. However, the effect of remediation in the third grade is still statistically distinguishable and of a meaningful magnitude as late as the seventh grade—five years after the retention decision.

That we continue to find a statistically significant and substantial treatment effect several years after the intervention despite the fade-out is very encouraging for the use of Florida-style test-based promotion policies, given that the effects of some other educational treatments have been found to fade completely over time. For instance, Puma et al. (2010) find that the initially positive effect of the Head Start program fades to the point of statistical insignificance by the end of first grade. The magnitude of the effect of treatment under Florida's test-based promotion policy five years later is comparable to that found for the five-year effect of assignment to small class sizes in elementary school (Nye, Hedges, and Konstantopoulos 1999).

The magnitude of the sustained effect of third-grade remediation under Florida's test-based promotion policy is noteworthy. By the time the second cohort subjected to the policy has entered the sixth grade, we continue to see a 0.183 standard deviation improvement in reading and a 0.174 standard deviation improvement in math due to remediation in the third grade. To put the size of that result into context, a 1 standard deviation in the quality of a child's teacher has been estimated to improve student achievement by at least 0.11 standard deviations the following year (Rivkin, Hanushek, and Kain 2005).

Given that policies can differ across school systems, it is important to note that our results are strictly related to test-based promotion policies identical in structure to Florida's program. We are not able to completely disaggregate the effect of retention from that of summer school attendance or teacher assignments during the retained year. However, we do provide some evidence that the policy's requirement that a student be assigned to a high-quality teacher the following year does not appear to drive the effects from treatment.

Though the results of this study are generally positive for the use of Florida's test-based retention policy, future research on this and similar programs is necessary. In particular, we are interested in the long-term outcomes of early grade retention in the form of the likelihood that a student graduates from high school as well as life outcomes such as labor market earnings. Further, given that it is relatively expensive to educate a child for an extra year, additional research is needed in order to determine whether the retention intervention is a cost-effective strategy for increasing student achievement in the long term.

We would like to thank the Florida K-20 Data Warehouse for providing the data necessary for the analysis. We would also like to thank Martin West for valuable comments.

REFERENCES

Alexander, Karl, Doris R. Entwisle, and Susan L. Dauber. 1994. O*n the success of failure: A reassessment of the effects of retention in the primary grades*. New York: Cambridge University Press.

Allen, Chiharu S., Qi Chen, Victor L. Wilson, and Jan N. Hughes. 2009. Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis* 31(4): 480–99.

Babcock, Philip, and Kelly Bedard. 2011. The wages of failure: New evidence on school retention and long-run outcomes. *Education Finance and Policy* 6(3): 293–322.

Greene, Jay P., and Marcus A. Winters. 2007. Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy* 2(4): 319–40.

Greene, Jay P., and Marcus A. Winters. 2009. The effects of exemptions to Florida's test-based promotion policy: Who is retained? Who benefits academically? *Economics of Education Review* 28(1): 135–42.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics* 18(5): 527–44.

Holmes, C. Thomas. 1989. Grade-level retention effects: A meta-analysis of research studies. In *Flunking grades: Research and policies on retention*, edited by Lorrie A. Shepard and Mary Lee Smith, pp. 16–33. Bristol, PA: Falmer Press.

Hughes, Jan N., Qi Chen, Felix Thoemmes, and Oi-man Kwok. 2010. An investigation of the relationship between retention in first grade and performance on high stakes tests in third grade. *Educational Evaluation and Policy Analysis* 32(2): 166–82.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2): 615–35.

Jacob, Brian A., and Lars Lefgren. 2004. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics* 86(1): 226–44.

Jacob, Brian A., and Lars Lefgren. 2007. The effect of grade retention on high school completion. NBER Working Paper No. 13514.

Jimerson, Shane R. 2001. A synthesis of grade retention research: Looking backward and moving forward. *California School Psychologist* 6: 47–59.

Jimerson, Shane R., Elizabeth Carlson, Monique Rotert, Byron Egeland, and L. Alan Sroufe. 1997. A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology* 35(1): 3–25.

Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. 1999. The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis* 21(2): 127–42.

Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid. 2010. *Head Start impact study: Final report*. Washington, DC: U.S. Department of Health and Human Services.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–58.

Roderick, Melissa, and Jenny Nagaoka. 2005. Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis* 27(4): 309–40.

Rust, James, and Kahryn A. Wallace. 1993. Effects of grade level retention for four years. *Journal of Instructional Psychology* 20(2): 162–66.

Schochet, P., T. Cook, J. Deke, G. Imbens, J. R. Lockwood, J. Porter, and J. Smith. 2010. *Standards for regression discontinuity designs*. Institute of Education Sciences. Available ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf. Accessed 17 January 2012.

Van der Klaauw, Wilbert. 2002. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review* 43(4): 1249–87.