REVISITING GRADE RETENTION: AN EVALUATION OF FLORIDA'S TEST-BASED PROMOTION POLICY

Jay P. Greene
Department of Education Reform
University of Arkansas
201 Graduate Education Building
Fayetteville, AR 72701

Marcus A. Winters
(corresponding author)
Department of Economics
University of Arkansas
201 Graduate Education Building
Fayetteville, AR 72701
winters@uark.edu

Abstract

In 2002, Florida adopted a test-based promotion policy in the third grade in an attempt to end social promotion. Similar policies are currently operating in Texas, New York City, and Chicago and affect at least 17 percent of public school students nationwide. Using individual-level data on the universe of public school students in Florida, we analyze the impact of grade retention on student proficiency in reading one and two years after the retention decision. We use an instrumental variable (IV) approach made available by the relatively objective nature of Florida's policy. Our findings suggest that retained students slightly outperformed socially promoted students in reading in the first year after retention, and these gains increased substantially in the second year. Results were robust across two distinct IV comparisons: an across-year approach comparing students who were essentially separated by the year in which they happened to have been born, and a regression discontinuity design.

1. INTRODUCTION

Several large public school systems have recently adopted test-based promotion policies for students in particular grades. Under these policies, students are required to demonstrate a certain level of academic preparation on a standardized test before they can be promoted to the next grade. There are usually various exemptions and alternative routes to promotion, but the default outcome under test-based promotion is that students with low test results are retained in the same grade. The public school systems of Florida, Texas, New York City, and Chicago now require all students in certain grades to achieve a minimal level on a standardized reading test in order to earn promotion to the next grade level. If such policies were to remain in only these school systems, they would affect more than 17 percent of the nation's public school students.[1]

While having students repeat a grade has long been a fairly common practice, the prevailing view among educators has been that it is in the best academic and social interests of students to advance to the next grade. In particular, teachers often believe that retaining a student will harm his or her self-esteem (Tompchin and Impara 1992; Jacob, Stone, and Roderick 2004). When students have been retained, it has generally been at the discretion of teachers in consultation with administrators and parents, and not based on the results of standardized tests.

Retaining a large population of students has direct and substantial economic consequences. If a student successfully completes his or her schooling, that student will spend (at least) an additional year in the public school system at an average cost to the taxpayer of $9,941 per year.[2] Further, by affecting academic growth and increasing the age at which the child will complete high school, these policies similarly affect the skills with which students enter the workforce and the probability that they will earn a high school diploma, which in turn is related to income (Hungerford and Solon 1987; Belman and Heywood 1991, 1997; Park 1994; Jaeger and Page 1996).

Proponents of these policies argue that the material taught in a given grade often assumes the student has a prerequisite knowledge base that was obtained in the prior grade. Thus, such students might benefit from shoring up their knowledge of basic material before they attempt the more difficult next grade. It is also possible that any such effect could grow over time as material becomes more and more difficult in later grades.

However, a wide body of existing empirical research suggests that retention in fact harms later academic progress (Peterson, DeGracie, and Ayabe 1987;

1. Author calculations using data from National Center for Education Statistics, *Digest of Education Statistics 2005*, Tables 33, 90.
2. National Center for Education Statistics, *Digest of Education Statistics 2005*, Table 162.

Holmes 1989; Alexander, Entwisle, and Dauber 1994; Jimerson et al. 1997; Roderick and Nagaoka 2005) and increases rates of students dropping out of school (Grissom and Shepard 1989; Roderick 1994; Jimerson 2001a, 2001b; Allensworth 2005). But the vast majority of this research is dated and focused on more subjectively guided retention rather than that driven by standardized testing.

By significantly reducing the subjectivity of retention decisions, test-based retention policies allow for an instrumental variable approach that was not available for previous researchers. These policies allow for substantial improvements on previous work, which demands a reopening of the empirical literature. Two previous studies (Jacob and Lefgren 2004; Roderick and Nagaoka 2005) have used the existence of test-based retention in Chicago to evaluate the impact of retention in that city.

This study adds to the limited research on the effects of test-based retention policies on student proficiency. We use a rich data set containing individual information on public school students in the state of Florida from 2002 to 2005. We use an instrumental variable approach with two distinct comparison groups to evaluate the robustness of any findings. In an across-year approach, we compare the gains of students who were retained during the first year of the policy with students who achieved the same low test scores but were not retained because they entered the third grade in the year prior and thus were not subject to retention. We then use a regression discontinuity approach similar to that used in the recent evaluations of Chicago's test-based retention policy in which we compare the performance of students just above and below the cutoff for retention during the third-grade year.

The results of these analyses show that retained students in Florida made significant and economically substantial reading gains relative to the control group of socially promoted students two years after being subjected to the policy. These benefits in reading from being retained grew substantially from the first to the second year after retention.

The remainder of this article proceeds as follows: In the next section we provide a brief overview of previous research on grade retention. We briefly describe Florida's policy and then discuss each of our comparison strategies and report their results independently. Finally, we conclude with a brief summary of our results and their implications for future research.

2. PREVIOUS RESEARCH

Several previous studies have evaluated the impact of grade retention on later student performance prior to the adoption of test-based promotion policies (Peterson, DeGracie, and Ayabe 1987; Holmes 1989; Alexander, Entwisle, and

Dauber 1994; Jimerson et al. 1997). Other researchers have focused on the effect of such policies on the probability that a student graduates from high school (Grissom and Shepard 1989; House 1998; Jimerson 2001a, 2001b; Roderick 1994; Allensworth 2005). Meta-analyses indicate that the cumulative finding of this previous research is that retaining a student leads to substantial academic harm (Holmes and Matthews 1984; Holmes 1989; Jimerson 2001a).

However, even relatively recent work on retention was limited by the difficulty of creating adequate groups with which to compare retained students. Lacking an objective policy, the decision to retain a student is a subjective judgment made by the teacher or school administrator. Thus most previous researchers focused on developing instruments based on observed characteristics. In his review of the research, Jimerson (2001a) reports that of twenty studies of grade retention in the 1990s, researchers most commonly matched samples based on some combination of IQ, socio-emotional adjustment, socioeconomic status (SES), or gender.

Although these observed characteristics were the best available for past researchers given the subjective nature of retention, their usefulness as instruments for retention are theoretically limited. While they clearly affect the probability that a student is retained, it is difficult to argue that any of the above characteristics are independent of test score gains. Moreover, even if the matched promoted students have test scores, IQs, SES, etc., similar to those of retained students, the fact that their teachers made opposite decisions about their promotion implies they may be dissimilar to retained students in some important way observed by the teacher but unobserved by the researcher. Thus, though often cited as conclusive, there is legitimate reason to doubt the usefulness of prior studies on subjective grade retention, and further research using exogenous instruments for retention is necessary to test the robustness of these previous results.[3]

The existence of more objective retention policies across the nation now provides researchers with an opportunity to create more meaningful groups with which to compare retained students than were available to researchers previously. Under the test-based promotion policies, students are much more likely to be retained if their score on a standardized test is below a certain threshold. Where previous retention decisions were endogenous to a host of unobservable factors, researchers can now use student performance relative to these arbitrarily set but objective cutoff scores as an exogenous instrument evaluating student performance under retention.

3. Roderick and Nagaoka (2005) provide a more in-depth review of this literature and come to a similar conclusion about its drawbacks. We defer to their review, given its recentness and our agreement with their conclusions.

Jacob and Lefgren (2004) and Roderick and Nagaoka (2005) used a regression discontinuity approach to study the impact of test-based promotion in Chicago. Both articles compared the gains of students whose test scores in the gateway grade were just above the threshold (most of whom were promoted) with those of students whose gateway test scores were just below the threshold (most of whom were retained). Both articles found that retained students made gains in reading and math after one year, but in the second year of the policy these gains disappeared in the third grade and were insignificant to negative in the sixth grade.

As in Florida, under the policy in Chicago students subject to retention were also required to attend summer school. In their analysis, Jacob and Lefgren are able to disentangle the impact of summer school and retention by taking advantage of the fact that in Chicago summer school students are given a chance to retake the exam and will avoid retention if they meet the test score threshold. This creates a new discontinuity separating those who received only the summer school treatment from those who received the summer school and retention treatments. In the third grade, their analysis found that after two years for third-grade students, summer school substantially increased average reading achievement, and retention had no impact on student achievement.

But the results of the research in Chicago may not generalize to all test-based promotion policies in other school systems. While both Florida's and Chicago's programs use test-based promotion, differences in the characteristics of the two programs could lead the policies to have different effects. For example, the Chicago program did not have a clear policy permitting exemptions to test-based promotion requirements, while Florida's did. Perhaps the restricted but guided discretion of educators' decisions about retention under Florida's test-based policy has significant advantages over the unguided policy in Chicago. In addition, recent allegations of testing impropriety in Chicago (Jacob and Levitt 2003) compared with validation of testing integrity in Florida (Greene, Winters, and Forster 2004; West and Peterson 2005) may produce different findings from the Chicago and Florida programs. If Chicago schools are manipulating test results in response to student retention rather than addressing the needs of those students, test-based retention may indeed be counterproductive. The current article contributes to the previous research by evaluating student performance one and two years after retention in Florida, using both across-year and discontinuity research designs.

3. FLORIDA'S TEST-BASED PROMOTION POLICY

In 2002 Florida began requiring third-grade students to meet at least the Level 2 benchmark (the second lowest of five levels) on the Florida Comprehensive Assessment Test (FCAT) reading test in order to be promoted

Table 1. Promotion Characteristics: All Students in Third Grade in 2002–3 with Scores below Test Score Threshold

Exemption for:	Percent
No code listed	3
Limited English proficient	7
Disability—testing not appropriate	<1
Passed alternative test	7
Student portfolio	3
Disability—has received extensive instruction	8
Already retained twice	1
No longer enrolled in school system	3
Academically promoted	12
Retained	59

Note: Percentages may not sum to 100 due to rounding.

to the fourth grade. According to the state's testing Web site, students who score at Level 2 are considered to have "limited success" with the challenging content on the test.[4] The entering third-grade class of 2002–3 was the first to be subjected to the mandate.

The legislature allowed for several exemptions to the retention policy. Exemptions were available to limited-English-proficiency students who had less than two years of instruction in English; disabled students whose individual educational plan indicated that testing was inappropriate for them; students who scored above the 51st percentile on another standardized reading test; students who were disabled and received intensive remediation in reading; students who demonstrated proficiency through a student portfolio; or students who had been retained twice previously.

Table 1 shows the promotion characteristics of third-grade students in the first year the policy was in place whose test scores were below Level 2 and for whom baseline test scores were reported in our data set. Only 59 percent of students who were subject to the policy and had test scores below the necessary threshold were actually retained in the third grade. The table shows that some students in the data set with scores below the threshold were coded as having been academically promoted without receiving an exemption. After discussions about this with the Florida data warehouse, it remains unclear why these students were promoted or whether there were errors in their coding.

4. Florida Department of Education, "FCAT Explorer: Parent and Family Guide." Available www.fcatexplorer.com/parent/shared/en/about_fcat.asp.

Due to this data anomaly, in the analyses that follow we identify students as retained from changes in the grade level of the test they were administered in each year rather than the state's classification.

Low-performing students who fail to receive an exemption from the policy are subject to retention as well as a series of other interventions. Retained students must be assigned to a "high-performing teacher" as determined by student performance data and above-satisfactory performance appraisals. Schools must develop academic improvement plans for these students that address their specific academic needs and create "success-based intervention strategies" for improvement, and they must create performance portfolios for the students. Retained students who fail to meet the necessary test score cutoff are required to attend a summer reading camp where they receive literacy instruction; during their retained year, the students must receive an additional ninety minutes of daily reading instruction.[5]

The existence of treatments other than retention suggests that any results from our estimation could be explained by the impact either of retention or of these other interventions, or some combination of the two. Unfortunately, in the analyses that follow we are unable to disentangle the impact of retention from that of these other reforms as Jacob and Lefgren did in Chicago. One reason for this failure is that the existence of alternative exemptions other than performance on an alternative test does not provide us with a way to identify a new discontinuity in who is retained or promoted. We are also unable to follow Jacob and Lefgren's procedure for those who passed an alternative test because the retesting procedure in Florida is less uniform than that of Chicago. Students may take the alternative exam at any point at least six weeks after taking the original test. Also, in order to receive promotion, Florida students must meet a much higher threshold on the alternative test than was required when the test was originally administered.

Since we are unable to disentangle these effects, our results must be thought of as an average treatment effect across each of the interventions of the Florida policy. Thus, for comparison, our results are more similar to the average treatment results of Jacob and Lefgren (2004) and the results of Roderick and Nagaoka (2005). We discuss the potential for explanations of the results other than the treatment of retention later in the article. However, throughout this article we refer to the treatment under Florida's policy as "retention," if only for the sake of brevity.

The only substantial change to Florida's retention policy since its implementation is that beginning in the 2004–5 school year, retained students

5. Florida Department of Education, *Third grade student progression*, available http://bsi.fsu.edu/schoolimprove/studentprogression/thirdgradeprog.htm

became eligible to receive a midyear promotion if they demonstrate possession of necessary skills. In the year evaluated in this article, retained students remained in the third grade for the entirety of the retained year.

Data

We utilize a rich data set, provided by the Florida Department of Education, containing test scores and demographic characteristics for the universe of students enrolled in grades 3–10 in a Florida public school from 2001–2 to 2004–5. The data set also includes a unique identifier for each child, allowing us to track the performance of individuals over time.

We analyze student-level test scores on the FCAT reading assessment, which the state administered to all students in grades 3–10 during each year under our analysis. Scores on the FCAT reading test are reported as developmental scale scores (DSS), which the Florida Department of Education designed to have the same meaning for proficiency across grades and years. That is, a student who scored 1,000 on the third-grade FCAT reading test in 2002 is thought to have identical absolute reading knowledge as a student who scored 1,000 on the fifth-grade FCAT reading test in 2004.

Our data set contained several instances of duplications of students, and most often the values on one or several of the variables were inconsistent across observations. After discussion with the Florida Department of Education, the reason for these duplicated instances of what should be a unique student identifier remains unclear. In cases in which all information for the duplicated cases was identical, we deleted all but one observation. In cases in which information differed, we excluded all observations of the student.

General Econometric Approach

We use an instrumental variables approach to evaluate treatment effects. We are interested in evaluating the impact of grade retention on student performance one and two years after the student's first third-grade year. That is, we are interested in the impact of the retention treatment on the treated so that we can then make inferences about the impact that treatment would have on those who would not have been treated before.

The structural equation treats a student's reading proficiency (Y) as a function of past reading proficiency, observable characteristics (Student), characteristics that are unobserved by the researcher (τ), whether the child received the treatment of the test-based promotion policy (retention and other treatments of the policy) (Z), and a stochastic error term (ε):

$$Y_{i,t+j} = \varphi_0 + \varphi_1 Y_{i,t} + \varphi_2 Student_i + \varphi_2 Z_{i,t} + \tau_i + \varepsilon_{i,t+j}, \quad (1)$$

where i indexes the student, t indexes time, and $j = 1, 2$ indexes the years since baseline.

A problem arises because whether the student receives the treatment (Z) is a function of unobserved factors that also affect his or her test score achievement:

$$Z_{i,t} = \beta_0 + \beta_1 Y_{i,t} + \beta_3 Student_i + \tau_i + \upsilon_i \quad (2)$$

where υ is a stochastic error term for the determination of whether a student is retained and all other variables are as previously defined.

In our case of studying grade retention, we worry that students with identical observed characteristics who are retained likely have a fundamental difference in their unobserved characteristics compared with those students who were promoted. Using Z as an explanatory variable will thus lead ordinary least squares (OLS) to be an inconsistent estimator of equation 1. This is the problem that has plagued previous research on retention based primarily on the subjective decisions of teachers.

In the context of regression discontinuity, Cameron and Trivedi (2005) discuss a two-stage procedure that produces consistent estimation of equation 1. In the first stage, we estimate the probability that a student is retained given the observable characteristics and value for the instrumental variable:

$$E[Z_{i,t}|Y_{i,t}, Student_i, \lambda_{i,t}] = \delta_0 + \delta_1 Y_t + \delta_2 Student_i + \delta_3 \lambda_{i,t} + \upsilon_i. \quad (3)$$

In the second stage, we can then estimate an equation similar to equation 1, but removing the unobserved portion and replacing $Z_{i,t}$ with $E[Z_{i,t}|Y_{i,t}, \text{Student}_i, \lambda_{i,t}] = \Pr[Z_{i,t} = 1|Y_{i,t}, \text{Student}_i, \lambda_{i,t}]$:

$$Y_{i,t+j} = \phi_0 + \phi_1 Y_{i,t} + \phi_2 Student_i + \phi_3 \Pr[Z_{i,t} = 1|Y_{i,t}, Student_i, \lambda_{i,t}] + \varepsilon_i, \quad (4)$$

where λ is an instrumental variable that is correlated with Z but is uncorrelated with ε. Under the regular assumptions, OLS produces consistent estimation of equation 4.

The normal procedure for solving such problems is to use instrumental variable methods (IV) such as two-stage least squares (2SLS). The classical 2SLS procedure would estimate both equations 3 and 4 using OLS. However, in our case the dependent variable in equation 3 is dichotomous, making OLS inappropriate. Probit, however, will produce a consistent estimate for equation 3. Thus in the analyses that follow we estimate equation 3 using probit to find $\Pr[Z_{i,t} = 1|Y_{i,t}, \text{Student}_i, \lambda_{i,t}]$ and then estimate equation 4 using OLS.[6]

6. We use the "predict" post-estimation command in STATA to produce $\Pr[Z_{i,t} = 1|Y_{i,t}, \text{Student}_i, \lambda_{i,t}]$.

We are not aware of a statistical software with a routine to directly produce these estimates with standard errors that are corrected for the two-stage nature of the model. Because of this we find standard errors with the bootstrapping technique using 1,000 repetitions. We also assume that ε is clustered by school in the calculation of these errors.

We pursue two different IV strategies that utilize different instruments (λ) in order to test the robustness of our findings. In one analysis, the across-year approach, we utilize the exogenous instrument of whether the student happened to have been in the third grade in the year the retention policy was implemented rather than the year before. In our second model, the regression discontinuity approach, we use whether the child's test score fell above or below the required test score for promotion within a very narrow neighborhood of the benchmark.

4. ACROSS-YEAR APPROACH

In our first analysis, we focus only on Florida students in the third grade in 2001–2 or 2002–3 whose test scores were below the Level 2 benchmark on the FCAT reading test. The score required to reach Level 2 was identical in both years.[7]

We compare the academic achievement of students with these low test scores who were in the first third-grade class (subject to the retention mandate) with the test score gains of students with the same low baseline scores but who entered the third grade in the year prior to the policy (who were thus not subjected to the program). On average, the two groups should be very similar, and any observed differences can be controlled statistically.

We estimate equation 3 using a binary variable indicating whether the students entered the third grade in the year before or after the policy's adoption as the instrument for retention (λ). In order to evaluate whether the policy's effects on student proficiency change over time, we compare the test score gains in reading in the first and second years after their initial third-grade year. For clarity, table 2 illustrates the grade and year comparisons made by our two evaluations. As the table shows, in the evaluation of gains after one year we compare the gains the control group made between 2001–2 and 2002–3 with the gains made by the treatment group between 2002–3 and 2003–4. Further, in the analysis of gains in the second year after retention, we compare the gains that the comparison group made between 2001–2 and 2003–4 with the gains that the treatment group made between 2002–3 and 2004–5.

7. Students were required to reach a score of 1,046 DSS points on the third-grade FCAT reading assessment.

Table 2. Across-Year Test Score Comparisons

	Control Group	Treatment Group
Baseline	2001–2 All students in third grade	2002–3 All students in third grade
Year 1	2002–3 Most students in fourth grade	2003–4 Most students in third grade
Year 2	2003–4 Most students in fifth grade	2004–5 Most students in fourth grade

The test scores of students in our two comparison groups differ not only in the year of the evaluation but in most cases in the grades evaluated as well. Since most students in the treatment group were retained after their baseline year, in the second year after baseline (2004–5) most of them were in the fourth grade. However, since they were not subjected to the retention policy, most of the students in the control group were initially promoted; thus in the second year after baseline (2003–4), most of them were in the fifth grade.

The existence of DSS scores allows us to make these comparisons across grades and years. However, some may worry about the ability of any test to produce accurate scores that are comparable across grade levels. Unfortunately, we are unable to use grade-level indicator variables in the analyses because whether the student was retained in the baseline year is completely collinear with grade level after one year and very strongly collinear with grade level in the second year. Past researchers in Chicago also did not control for student grade level, presumably for the same reason. Thus we are forced to rely on the validity of the DSS scores in this analysis.[8]

The benefit of the across-years comparison group is that it utilizes observations for all students who were retained by the policy, not just a limited cohort of those near the passing threshold as in our second strategy and that of the previous research in Chicago. We might worry that the treatment's impact could be nonlinear across student baseline proficiency levels. If this were the case then the across-year strategy could allow for a more comprehensive evaluation of the effects of the retention policy than a regression discontinuity design.

One important limitation of the across-years approach, however, is that the treatment and control groups do not have identical demographic characteristics. Table 3 reports descriptive statistics on the treatment and control groups

8. Those interested in how the DSS scores were produced should refer to a memorandum from Deputy Commissioner Betty Coxe to Florida school district superintendents on 14 August 2002, with the subject "FCAT Developmental Score Scale." This memorandum includes technical information on the development and validity of these scores. Available http://info.fldoe.org/docushare/dsweb/Get/Document-473/DPSM03-015/pdf.

Table 3. Summary Statistics for Variables Used in Across-Year Analyses

	YEAR 1 ANALYSIS			YEAR 2 ANALYSIS		
	All Students	**Control Group**	**Treatment Group**	**All Students**	**Control Group**	**Treatment Group**
Indian	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%
Asian	1.0%	1.1%	0.9%	1.0%	1.1%	0.9%*
African American	37.3%	37.7%	36.7%*	37.5%	37.9%	36.9%*
Hispanic	29.2%	28.0%	31.0%*	29.4%	28.1%	31.1%*
Multiple race	1.8%	1.7%	1.8%	1.8%	1.7%	1.8%
White, not Hispanic	30.5%	31.3%	29.4%*	30.2%	30.9%	29.1%*
Free or reduced price lunch	74.6%	73.2%	76.6%*	74.8%	73.4%	76.7%*
Baseline math	1014.84	1021.03	1005.98*	1015.50	1022.47	1005.77*
Baseline reading	770.24	764.79	778.04*	770.98	765.86	778.12*
Year 1 reading	1130.52	1126.77	1135.90*			
Year 2 reading				1135.72	1225.25	1298.28*
Retained (baseline)	28.6%	6.0%	60.9%*	27.7%	5.9%	58.2%*
N	73,695	30,299	43,396	71,950	41,917	30,033

*statistically significant at 5%.

for variables used in the regressions and compares them using a Kruskal-Wallis test. The table shows that the two groups of students are, in fact, statistically different on most observed dimensions. The descriptive statistics are slightly different for the two analyses because of student migration out of Florida or because the test score information in the second year was unobserved for some other reason.

It is worth noting that although each of the demographic differences is statistically significant, most are arguably insubstantial. Thus the primary demographic difference between the two cohorts remains the year in which they happened to have been born, which alone determined whether they were subject to Florida's retention policy. However, that these demographic differences are statistically significant strongly suggests the use of statistical controls where available and use of an alternative comparison strategy to check for robustness of any results.

Table 4 reports descriptive information on reading test scores disaggregated by whether or not the student was actually retained. Note that after two years on average, retained students appear to have made substantial improvements in reading proficiency compared with promoted students. However, though these descriptive statistics are suggestive, we need to estimate the models in order to properly identify any relationship.

Table 4. Descriptive Information on Reading Test Scores by whether Students Were Retained

	ALL STUDENTS		RETAINED		PROMOTED	
	Mean	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Baseline						
Reading	770.24	257.23	762.41	246.60	773.37	261.30
Year 1 reading	1130.52	318.04	1102.47	307.44	1141.76	321.50
Year 2 reading	1255.73	303.15	1346.85	293.13	1220.81	299.66

The results of our analyses for the across-year comparison on the test score gains made in reading are reported in table 5. In both the first- and second-year evaluations, the coefficient on "retained"—the variable indicating the probability that the child was retained—is statistically significant and positive, indicating that retained students outperformed promoted students in both years. However, the size of the coefficient for retention increases substantially from the first year after the student was retained (20.34), when treated students were completing their second third-grade year, to the second year after the students were retained (152.10), when most of the treated students were in the fourth grade (recall table 2 for clarity in the comparisons). These results indicate that the reading benefit of retention in the third grade increased in the second year, when most of the retained students were in the fourth grade.

We can put the above results into a more manageable context by converting them into standard deviations from the third-grade reading test score. In 2002–3, the baseline third-grade year for our treatment group, the mean DSS score on the FCAT reading test for all students was 1290.9 with a standard deviation of 381.2. Thus after one year we find that retained students outperformed promoted students by about 0.05 standard deviations. The reading benefit of retention after two years was an economically substantial 0.40 standard deviations.

Limitations of Across-Year Approach

The across-year comparison analyzed in the previous section provides a meaningful estimate of the impact of grade retention on the academic gains of low-performing students. However, this approach is limited in a few important ways that deserve mention.

First, as described above, the treatment and control groups in the across-year comparison differ in their descriptive statistics. Though these differences are quite small, they are nonetheless worrisome, and could indicate that students differ to some extent in unobserved factors as well. Controlling for the

Table 5. Regression Results: Across-Year Comparison

	ONE YEAR			TWO YEARS		
Variable	**Coefficient**	**Standard Error**	**P-Value**	**Coefficient**	**Standard Error**	**P-Value**
Constant	505.33	6.66	0.00*	699.90	6.45	0.00*
Baseline reading	0.39	0.00	0.00*	0.32	0.01	0.00*
Baseline math	0.35	0.00	0.00*	0.29	0.00	0.00*
Indian	−0.65	22.24	0.98	19.06	20.01	0.34
Asian	38.33	9.52	0.00*	62.82	9.41	0.00*
African American	−9.70	3.09	0.00*	−19.06	3.20	0.00*
Hispanic	8.87	3.51	0.01*	22.85	3.50	0.00*
Multiple race	30.28	7.18	0.00*	24.26	7.27	0.00*
Free or reduced price lunch	−43.45	2.82	0.00*	−32.14	2.76	0.00*
Retained	20.34	5.27	0.00*	152.10	6.23	0.00*
Adjusted R^2	0.2588			0.2002		
N	73,695			71,950		

Notes: The dependent variable is the student's reading score in the baseline year. Standard errors are produced through bootstrapping and are clustered by school. "Retained" is the probability that a student is retained given observed characteristics and whether the student was in the third grade in the first year the policy was implemented, which is the result from estimation of a first-stage probit equation. Only results of the second stage are shown here. *statistically significant at 5%.

observed factors with binary variables indicating student race and free lunch classification status helps to limit any bias, though not to overcome it completely.

The across-year comparison approach is also limited because our treatment and control groups entered the third grade in different years. It is possible that students in our treatment and control groups were not uniformly affected by reforms other than the retention policy that might have occurred in Florida. In fact, Florida is the leader in many educational reforms. In 1998 Florida began a voucher program for students in chronically failing schools. The state has also seen strong growth in the use of charter schools during this period. Finally, Florida began to implement the Reading First program in certain elementary schools during this time (beginning with a group of schools in 2003–4). Our results could be biased if our treatment and control groups were affected by these other policies in different ways.

Further, it is possible that schools responded to the implementation of the retention policy by improving the education provided in the third grade so that fewer students would be retained. That is, there could be a pre-treatment effect that raised student proficiency among those at the bottom of the distribution

that could confound our results. The higher baseline reading scores for our treatment group reported in table 4 indicate that this bias likely exists.

Finally, it is possible that our estimates could be biased by any impact on academic performance caused by differences in the peers of retained and promoted students after baseline. In particular, because of the high retention rates caused by the policy, those students retained under the policy will now attend classrooms with a higher percentage of students who were retained (and are clearly low performing) than promoted students. If there are peer effects on student ability, then we are failing to account for an important difference in the academic experience of these two groups of students. It is worth noting, however, that this reasoning suggests that the academic performance of treated students should be brought down by their less academically advanced classmates; thus if there are peer effects at work, our already substantially positive results for the treatment variable are likely biased downward.

Regression Discontinuity Comparison

Due to the limitations of the across-year approach, we further analyze the effect of Florida's retention policy using a regression discontinuity design. This design also allows us to more directly compare our results with those of the recent analyses of Chicago's test-based retention policy. We stress, however, we are not directly estimating the model used in these previous analyses. One of the more important differences is that we very likely draw the discontinuity measure at a different point and are unable to control for test scores before the third-grade year. We are also unable to disaggregate the impact of retention from that of other coinciding treatments, as were Jacob and Lefgren in Chicago. However, to evaluate the average treatment effect, our design remains quite similar to that previous analysis and provides a robustness check for our across-year findings.

The use of regression discontinuity has been growing in popularity as a tool for evaluating public policy. This design is useful in cases, such as the present one, in which a treatment is primarily determined by the subject reaching a particular measurable threshold. Van der Klaauw (2002) shows that if obtaining a treatment is conditioned on meeting a certain known threshold, an analysis of individuals in a narrow margin around this threshold approximates random assignment.

We take advantage of the existence of the known cutoff score under Florida's policy, below which students were more likely to be retained and above which they were more likely to be promoted. Similar regression discontinuity designs have been utilized in other areas of economics of education (Angrist and Lavy 1999; Van der Klaauw 2002; Chay, McEwan, and Urquiola 2005) and in studies outside of education (see, e.g., DiNardo and Lee 2004).

Table 6. Summary Statistics for Variables Used in Regression Discontinuity Analyses

	All Students	Control Group	Treatment Group
Indian	0.3%	0.3%	0.4%
Asian	1.2%	1.2%	1.1%
African American	36.2%	36.7%	35.8%
Hispanic	26.0%	25.9%	26.2%
Multiple race	2.3%	2.3%	2.3%
White, not Hispanic	33.9%	33.5%	34.3%
Free or reduced price lunch	75.1%	75.6%	74.7%
Baseline math	1006.86	1009.69	1004.30
Baseline reading	1046.12	1060.32	1033.25*
Year 1 reading	1357.77	1356.61	1358.82
Year 2 reading	1427.38	1392.11	1459.36*
Retained (baseline)	25.8%	4.7%	45.0%*
N	7,087	3,370	3,717

*statistically significant at 5%.

In this evaluation we compare the test score gains of students whose reading score in 2002–3 was just below the threshold required for promotion to students who were in the third grade that same year and whose scores were just above this threshold. Unlike the previous analysis, all students in this design were in the third grade in 2002–3 and were vulnerable to the policy if they did not score above the necessary threshold.

We draw the neighborhood of "close" to the cutoff for promotion around those students whose score on the third-grade FCAT reading test in 2002–3 (the test used for the retention decision) was within 25 DSS points of the threshold for promotion. The minimum DSS score required to avoid the retention policy was 1046, and in the baseline year the mean DSS score on the FCAT reading test for all students was 1290.9 with a standard deviation of 381.2.

The regression discontinuity design addresses most of the concerns raised about the across-year approach. First, table 6 shows that the observable characteristics for students in this narrow neighborhood around the test score benchmark are statistically identical, except, of course, for their baseline reading test score and whether or not they were retained. Unlike the across-year case, here we are able to observe information about all students in both the first and second years after baseline, so we do not need to produce different descriptive statistics for both equations.

The regression discontinuity design also addresses concerns that the results of the across-year comparison could be biased due to heterogeneity in

the academic experiences of students in the treated and control groups. In particular, because all students in the regression discontinuity comparison entered the third grade in the same year, they should be similarly affected by policies that were operating at the same time as the promotion policy. That students experienced their first third-grade year at the same time also suggests that our results should not be biased by any pre-treatment effect where schools could improve instruction in order to decrease the number of students who are retained under the policy. This is particularly true here since we are able to control for the student's math and reading proficiency in the baseline year.

Table 6 also reports the percentage of students in the treatment and control groups of the discontinuity approach who were retained and exempted from the policy. Under the twenty-five-point definition, the table shows that in fact 55 percent of students with scores below the test score cutoff were actually promoted. Thus the regression discontinuity in this article follows the so-called "fuzzy" design. That is, the discontinuity of student baseline test scores is not strict. The use of the fuzzy discontinuity design implies use of a two-stage approach to accurately measure the effect of the policy. As in the across-year comparison, we again adopt such a two-stage IV analysis to account for the exemptions to the policy.

The regression discontinuity approach also suffers from a potential problem with external validity not faced by our across-year approach. By limiting the analysis to only those students whose baseline score is within a quite narrow region of the cutoff score, we are only able to make inferences about the effect of the policy on this small group of marginally affected students. If the impact of the policy is nonlinear across baseline student ability, then students with very low baseline proficiency will be more or less affected by the policy and our estimates will not indicate the comprehensive effect of retention. We would have concerns that such nonlinearity in the treatment effect could be at work if the estimates from the regression discontinuity approach are substantially different from those of the across-year approach reported earlier.

We estimated equations 3 and 4 with bootstrapped standard errors clustered by school. In this analysis we use an indicator variable for whether or not the student's baseline reading score was above or below the threshold for retention as the instrumental variable, λ.

The results of the regression discontinuity comparison are consistent with those of the across-year approach. Table 7 reports the reading results after one and two years.

The first set of columns in table 7 shows that after one year, retained students made reading gains on the FCAT that were not statistically different from those made by promoted students. However, the second set of columns indicates that these relative gains grew to 176.90 DSS points in the second year

Table 7. Regression Results: Regression Discontinuity Comparison

	ONE YEAR			TWO YEARS		
Variable	**Coefficient**	**Standard Error**	**P-Value**	**Coefficient**	**Standard Error**	**P-Value**
Constant	919.64	346.08	0.01*	1230.85	389.23	0.00*
Baseline reading	0.44	0.32	0.17	0.18	0.37	0.63
Baseline math	0.00	0.01	0.80	0.00	0.01	0.73
Indian	5.48	29.43	0.85	15.83	30.15	0.60
Asian	29.58	16.66	0.08	61.19	21.34	0.00
African American	−43.50	5.93	0.00*	−59.12	6.46	0.00*
Hispanic	6.99	5.90	0.24	8.38	6.45	0.19
Multiple race	14.38	14.35	0.32	−1.51	16.49	0.93
Free or reduced price lunch	−9.62	2.30	0.00*	−10.78	2.72	0.00*
Retained	31.71	25.20	0.21	176.90	28.34	0.00*
Adjusted R^2	0.018			0.0451		
N	7,087			7,087		

Notes: The dependent variable is the student's reading score in the baseline year. Standard errors are produced through bootstrapping and are clustered by school. "Retained" is the probability that a student is retained given observed characteristics and whether the student had reading proficiency in the baseline year that was above the threshold for promotion under the policy. Only results of the second stage are shown here. *statistically significant at 5%.

after retention, and this result is statistically significant at any conventional level. In standard deviation terms, our results indicate that after two years, students who were retained outperformed promoted students by about 0.46 standard deviations in reading, which is larger but similar to the estimated impact using the across-year approach.

5. INTERPRETATION AND DISCUSSION

The results of both the across-year and regression discontinuity approaches suggest that students subjected to the treatment of Florida's test-based retention policy made significant and economically substantial gains in reading relative to promoted students. Further, that the impact of the policy for reading scores grows after two years is consistent with the idea that retained students will continue to gain ground in reading relative to promoted students in later years as academic material becomes more difficult. The fact that the size of the impact found after one and two years is quite similar across these two quite different comparison approaches provides confidence that our results are robust.

The results indicate that the impact of Florida's retention policy was to increase reading proficiency quite substantially after only two years. Further, when interpreting the size of this effect, it is important to keep in mind that retained students receive lower-level instruction than do socially promoted students in any given year. Thus our results from both the across-year and regression discontinuity comparisons indicate that retained students have about 0.40 standard deviations higher academic proficiency than promoted students even though promoted students are taught at a level where they should be able to reach a higher maximum test score gain. This is further demonstrated by looking back at the descriptive statistics reported in table 3. The mean reading score two years after baseline indicates that on average, those students who were retained entered the fifth grade with a higher DSS score than was possessed by promoted students when they left the fifth grade.

The analyses in this article add to the previous research in several important ways. First, the fact that our positive result contradicts past findings of large harms from retention in analyses of subjective retention policies is of great interest and is consistent with the work in Chicago. These differences can be explained by the benefits of a more objective retention policy for developing groups to which we can compare the performance of retained and promoted students. Our differing results indicate that researchers should focus on the new widespread availability of objective retention policies to reopen the empirical discovery of the effect of grade retention.

Also, until this analysis we had only evidence of a test-based retention policy over time from one school system using a similarly strong research design. Given the wide scope of these policies, their growing popularity among school systems, and variations in the policy designs, it is important that the evaluation of their effectiveness includes a large scope of areas in which they have been tried.

Future research is necessary in order to understand why the results from Florida are different from those from Chicago. One significant difference between our results and those of Jacob and Lefgren (2004) is our inability to separate out the impact of retention from that of other simultaneous treatments. In fact, in their evaluation of Chicago's similar policy, Jacob and Lefgren found that the impact of retention in third grade on reading scores goes to zero after two years, while there is still a positive impact from summer school after this time. However, the size of the impact found in our analyses of Florida is much larger than Jacob and Lefgren found from summer school, making it unlikely that summer school is driving the entirety of our results. Further, while in Chicago the overall impact of retention and summer school decreased to zero after two years, our results indicate that the largest impact of the reform is found after this period. Future research on similar policies in

other school systems is necessary in order to better understand the differences in the results from Florida and Chicago.

Future research is also necessary to discover the even longer-term relationship between early retention (specifically in grade three) and academic proficiency. Further, future research on the effect of test-based promotion on the probability that a student will graduate from high school is also necessary. We might expect that improving student proficiency would increase the probability that a student graduates. However, retention also increases the number of years that a student is in school and the age at which he or she is eligible to graduate. These factors could make dropping out of high school more attractive to a student than otherwise. At least some previous research suggests that earning a high school diploma itself improves income regardless of educational attainment (Hungerford and Solon 1987; Belman and Heywood 1991, 1997; Park 1994; Jaeger and Page 1996). Thus the overall impact of grade retention could be negative if it improves academic proficiency but increases the probability that a student drops out.

It is also important to emphasize that although the results of this analysis provide evidence that retention policy in the third grade improved academic proficiency in Florida, it is possible that policies in other grades could have other effects. Specifically, it is possible that the negative effect of retention on a student's self-esteem could increase and the positive impact of retention on a student's relative knowledge could decrease in later grades, as students become more attached to their peers and the difference in knowledge between grades changes. It is worth noting that the findings from Chicago do indicate that the effect from retention was more harmful in the sixth grade than in the third grade.

Although there is a need for an expansive additional research program evaluating the effects of grade retention on student performance, this study provides evidence that Florida's policy has substantially improved the academic proficiency of the lowest-performing students in the state. Our results are consistent using two very different comparison strategies, indicating a robust positive effect from test-based retention in Florida.

We wish to thank Bruce Dixon, Julie Trivitt, and an anonymous reviewer for their valuable comments and suggestions. We also thank the Manhattan Institute for Policy Research for their support of this work. All remaining errors, of course, are our own.

REFERENCES

Alexander, K. L., D. R. Entwisle, and S. L. Dauber. 1994. *On the success of failure: A reassessment of the effects of retention in the primary grades.* New York: Cambridge University Press.

Allensworth, E. M. 2005. Dropout rates after high-stakes testing in elementary school: A study of the contradictory effects of Chicago's efforts to end social promotion. *Educational Evaluation and Policy Analysis* 27 (4): 341–64.

Angrist, J. D., and V. Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114 (2): 533–75.

Belman, D., and J. S. Heywood. 1991. Sheepskin effects in the returns to education: An examination of women and minorities. *Review of Economics and Statistics* 73: 720–24.

Belman, D., and J. S. Heywood. 1997. Sheepskin effects by cohort: Implications of job matching in a signaling model. *Oxford Economic Papers* 49: 623–37.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Chay, K. Y., P. J. McEwan, and M. Urquiola. 2005. The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review* 95 (4): 1237–58.

DiNardo, J., and D. S. Lee. 2004. Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics* 119 (4): 1383–441.

Greene, J. P., M. A. Winters, and G. Forster. 2004. Testing high-stakes tests: Can we believe the results of accountability tests? *Teachers College Record* 106 (6): 1124–44.

Grissom, J. B., and L. A. Shepard. 1989. Repeating and dropping out of school. In *Flunking grades: Research and policies on retention*, edited by L. Shepard and M. Smith, pp. 34–63. London: Falmer Press.

Holmes, C. T. 1989. Grade-level retention effects: A meta-analysis of research studies. In *Flunking grades: Research and policies on retention*, edited by L. A. Shepard and M. L. Smith, pp. 28–33. London: Falmer Press.

Holmes, C. T., and K. Matthews. 1984. The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research* 54 (2): 225–36.

House, E. R. 1998. The predictable failure of Chicago's student retention program. Paper presented at the Conference on Rethinking Retention to Help All Students Succeed, Chicago, November. Available www.designsforchange.org/pdfs/house.pdf. Accessed 30 May 2007.

Hungerford, T., and G. Solon. 1987. Sheepskin effects in the returns to education. *Review of Economics and Statistics* 69: 175–77.

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

Jacob, B. J., and S. D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118 (3): 843–77.

Jacob, R. T., S. Stone, and M. Roderick. 2004. *Ending social promotion: The response of teachers and students*. Chicago: Consortium on Chicago School Research.

Jaeger, D. A., and M. E. Page. 1996. Degrees matter: New evidence on sheepskin effects in the returns to education. *Review of Economics and Statistics* 78: 733–40.

Jimerson, S. R. 2001a. Meta-analysis of grade retention research: Implications for practice in the twenty-first century. *School Psychology Review* 30 (3): 420–37.

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

Jimerson, S. R., E. Carlson, M. Rotert, B. Egeland, and L. A. Sroufe. 1997. A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology* 35 (1): 3–25.

Park, J. H. 1994. Estimation of sheepskin effects and returns to schooling using the old and the new CPS measures of educational attainment. Working paper, Princeton University.

Peterson, S. E., J. S. DeGracie, and C. R. Ayabe. 1987. A longitudinal study of the effects of retention/promotion on academic achievement. *American Educational Research Journal* 24 (1): 107–18.

Roderick, M. 1994. Grade retention and school dropout: Investigating the association. *American Educational Research Journal* 31 (4): 729–59.

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

Tompchin, F. M., and J. C. Impara. 1992. Unraveling teachers' belief about grade retention. *American Educational Research Journal* 29: 199–223.

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

West, M. R., and P. E. Peterson. 2005. The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. Paper presented at the Annual Conference of the Royal Economic Society, University of Nottingham, March.