Education Working Paper Archive

Getting Farther Ahead by Staying Behind:

A Second-Year Evaluation of Florida's Policy to End Social Promotion

September 14, 2006

Jay P. Greene
Endowed Chair and Head of the Department of Education Reform,
University of Arkansas

Marcus A. Winters
Doctoral Academy Fellow, Department of Education Reform,
University of Arkansas

Executive Summary

Social promotion has long been the normal practice in American schools. Critics of this practice, whereby students are promoted to the next grade regardless of academic preparation, have suggested that students would benefit academically if they were made to repeat a grade. Supporters of social promotion claim that retaining students (i.e, holding them back) disrupts them socially, producing greater academic harm than promotion would. A number of states and school districts, including Florida, Texas, Chicago, and New York City, have attempted to curtail social promotion, by requiring students to demonstrate academic preparation on a standardized test before they can be promoted to the next grade.

This study analyzes the effects of Florida's test-based promotion policy on student achievement two years after initial retention. It builds upon our previous evaluation of the policy in two ways. First, we examine whether the initial benefits of retention observed in the previous study continue, expand, or contract in the second year after students are retained. Second, we determine whether discrepancies between our evaluation and the evaluation of a test-based promotion policy in Chicago are caused by differences in how researchers examined the issue, or by differences in the nature of the programs.

Our analysis shows that, after two years of the policy, retained Florida students made significant reading gains relative to the control group of socially promoted students. These academic benefits grew substantially from the first to the second year after retention. That is, students lacking in basic skills who are socially promoted appear to fall

further and further behind over time, whereas retained students appear to be able to catch up on the skills they are lacking.

Further, we find these positive results in Florida, both when we use the same research design that we used in our previous study, and when we use a design similar to that employed by the evaluation of the program in Chicago, The differences between the Chicago and Florida evaluations appear to be caused by differences in the details of the programs, and not by differences in how the programs were evaluated.

Introduction

Social promotion is the practice of promoting students to the next grade regardless of their academic preparation. While some students have always been made to repeat a grade, 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. 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.

This practice of social promotion has recently been replaced by "test-based promotion" in a number of states and school districts around the country, including Florida, Texas, Chicago, and New York City. Under test-based promotion, 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.

There has been considerable debate among educators, policymakers, and researchers about the consequences of this shift away from social promotion and toward test-based promotion. This study adds evidence to that debate by analyzing the effects of being retained under Florida's test-based promotion policy on student achievement two years after initial retention. This study builds upon our previous evaluation of the policy in two ways. First, we are able to examine whether the initial benefits of retention under a test-based policy observed in the previous study continue, expand, or contract in the second year after students are retained. Second, we are able to determine whether the different findings of our evaluation and a high-quality evaluation of a test-based promotion policy in Chicago are caused by differences in how the researchers examined the issue or by differences in the nature of the programs.

The results of this new analysis show that retained students in Florida made significant reading gains relative to the control group of socially promoted students two years after being subjected to the policy. These academic benefits of being retained grew substantially from the first to the second year after retention. That is, students lacking in basic skills who are socially promoted appear to fall further and further behind over time, whereas retained students appear to be able to catch up on the skills they are lacking. In addition, we find these positive results for the test-based promotion policy in Florida whether we use the same research design that we used in our previous study or a design similar to that employed by the evaluation of the program in Chicago. The differences in outcomes from the Chicago and Florida evaluations appear to be caused by differences in the details of the programs and not by differences in how the programs were evaluated.

Previous Research on Discretionary Retention

Under the practice of social promotion, some students have always been retained, but retention was rare and was based on the discretion of educators, not the results of standardized tests. Several previous studies have evaluated the academic impact of this discretionary retention under social promotion regimes. 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 2001).

These findings on the effects of discretionary retention are plagued by two serious limitations. First, it is very hard for those studies to find an appropriate control group against which retained students could be compared. Even if control-group students have similar test scores and other observable characteristics, students retained at the discretion of educators may differ significantly in unobservable ways. When educators use their discretion to retain students, they are aware of detailed contextual information that may lead them to recommend retaining one student while promoting another student with similar test scores and other recorded characteristics.

The fact that educators *chose* to retain one student and not another means that the two are not likely to be similar in their future prospects. After all, if the two really had been identical, educators would probably have made the same decision about their retention. The retained students' unrecorded disadvantages may account for their lower future achievement, not their retention. Unfortunately, most of the previous studies used in the meta-analyses that draw negative conclusions about retention failed to address this difficulty with proper techniques or research design to produce valid apple-to-apple comparisons. While these meta-analyses are often cited as conclusive, there is legitimate reason to doubt the findings of previous studies on discretionary grade retention.[1]

Second, it is not at all clear that the findings from studies of discretionary retention under social promotion regimes would apply to retention under test-based promotion policies. Studies of discretionary retention are essentially evaluations of whether educators use their discretion wisely in identifying students who ought to be retained. If that discretion is used wisely, only students who could benefit from retention are retained and all others are promoted.

[1] Roderick and Nagaoka (2005) provide a very useful review of this literature and come to a similar conclusion.

Under test-based promotion policies, the discretion of educators is greatly restricted. Retention decisions are based primarily or exclusively on the results of standardized tests. This shift to test-based promotion has been motivated by the belief that educators have generally not used their discretion wisely, either by failing to retain more students or by failing to retain the right students. It would therefore be inappropriate to extrapolate from evaluations of discretionary retention to the effects of retention under test-based policies meant to restrict or alter the use of that discretion.

Previous Research on Test-Based Retention

In addition to our previous evaluation of Florida's test-based promotion policy (Greene and Winters 2006), there is another high-quality study of test-based retention.[2] Roderick and Nagaoka (2005) evaluated the impact of a test-based promotion policy in Chicago on reading-test scores. Since 1996, students in Chicago have been required to reach minimal benchmarks on the reading and math portions of the Iowa Test of Basic Skills (ITBS) in the third, sixth, and eighth grades in order to be promoted to the next grade. Roderick and Nagaoka found that the retention policy led to small improvements in reading scores relative to socially promoted students during the first year after the retention decision but that these gains disappeared or turned negative in the following year.

The existence of a test-based promotion policy in Chicago allowed Roderick and Nagaoka (2005) to develop more appropriate comparison groups than had been available to previous researchers. They utilized two comparison groups in the study. First, they

[2] The results reported in this paper after one year differ somewhat from those reported in Greene and Winters (2006) because of revisions to the original dataset obtained from the state of Florida as well as slightly different analytical models.

took advantage of a change in the policy's design that made it likely that students with scores just below the test-score cutoff would get an exemption and thus be promoted in a later year. Prior to this change, students with scores just below the cutoff were likely to be retained; after the change, students with these same scores were likely to be promoted. Roderick and Nagaoka (2005) compared the test-score gains of these two groups on the assumption that the only difference between them was the year in which the student happened to have been born. This was the "across-year" research design.

In a second comparison, Roderick and Nagaoka (2005) took advantage of the existence of an observable cutoff for the promotion policy and utilized a regression discontinuity design. In this design, they included only students with test scores that were very close but on either side of the cutoff score. That is, they compared the test-score gains of students whose original score was "just" above the necessary threshold (most of whom were promoted) with those of students in the same year whose score was "just" below the threshold (most of whom were retained). This was their "discontinuity" research design.

Using multiple analytical models on both the across-year and discontinuity research designs, Roderick and Nagaoka (2005) found similar results. They found that the retention policy in Chicago had a mild positive impact on the test-score performance of retained students relative to promoted students in the year that the students were retained. However, in their analysis of test scores two years after the baseline year, each specification found that the effect of retention was either statistically insignificant or negative.

But this negative result from Roderick and Nagaoka's study in Chicago may not be generalizable to all test-based promotion policies in other school systems. Perhaps Chicago's test-based promotion policy has been counterproductive while Florida's has been beneficial. While both 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 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 (see Jacob and Levitt 2003and Greene, Winters, and Forster 2002) compared with validation of testing integrity in Florida (see 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 paper analyzes student performance one and two years after retention in Florida, using both across-year and discontinuity research designs. If an analysis in Florida were to produce negative results, like those found by Roderick and Nagaoka (2005) in Chicago, we could have greater confidence that test-based retention policies truly harm student achievement. However, if the results differ even when similar analyses are performed, we have reason to be more encouraged about the prospects of test-based promotion as practiced and implemented in Florida. Especially given the clearer exemption policy and superior test integrity in Florida, a positive result from Florida in a second-year study using multiple research designs would suggest that test-based promotion is likely to add significantly to student learning under the proper conditions.

Florida's Test-Based Promotion Policy

In 2002, the Florida legislature voted to require third-grade students to meet at least the Level 2 benchmark (the second-lowest of five levels) on the FCAT reading test in order to be promoted to the fourth grade. According to the state's testing website, students who score at Level 2 are considered to have "limited success" with the challenging content on the test.[3] The third-grade class of 2002–03 was the first that was subjected to the mandate.

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

Table 1 shows the promotion characteristics of third-grade students in the first year that the policy was in place, whose test scores were below Level 2 and for whom baseline test scores were reported in our dataset. The table shows that only 57 percent of students who had test scores below the threshold necessary to be promoted were actually retained in the third grade. The table shows that some students (13 percent) with scores

[3] Florida Department of Education, "FCAT Explorer: Parent & Family Guide," http://www.fcatexplorer.com/parent/shared/en/about_fcat.asp.

below the threshold were coded as having been promoted without any explanation for their exemption. After discussing this with the Florida data-warehouse personnel, it remains unclear why these students were promoted or whether there was an error in their coding.[4]

Schools must develop an academic improvement plan for any student who does not meet the standards for promotion. These plans must address the student's specific academic needs and create "success-based intervention strategies" for his improvement.[5] Students who fail to meet the necessary test-score cutoff are also required to attend a summer reading camp, where they receive literacy instruction.

The only substantial change to Florida's retention policy since its implementation is that beginning in the 2004–05 school year, retained students became eligible to receive a midyear promotion if they demonstrate possession of necessary skills. In the time period evaluated in this paper, retained students remained in the third grade for the entirety of the retained year.

Research Design

The most difficult problem for previous studies evaluating the academic effect of grade retention has been the identification of a proper group with which to compare retained students. The existence of a test-based retention policy helps solve this problem by reducing (but not eliminating) the impact of subjective teacher assessments that made comparisons difficult in the past. With the increased reliance on objective, test-based criteria for promotion, we can identify treatment and control groups that are similar on

[4] E-mail exchange between authors and Florida K-20 data-warehouse representative, May 10, 2006.

[5] Florida Department of Education, "Promotion and Retention: Common Questions and Answers," http://www.firn.edu/doe/commhome/progress/promo-qa.pdf.

those criteria and are less likely to differ in other, unrecorded ways. We can also use more advanced econometric techniques to ensure the comparability of our comparison groups.

In this paper, we utilize two strategies for identifying comparison groups with which to evaluate the effect of grade retention. In the first analysis, we compare students with similar reading-test scores who differ by the year in which they entered the third grade. In the second analysis, we utilize the discontinuity in retention created by the test-score threshold and compare the achievement of students who were just above and just below the retention benchmark.

Across-Year Comparison

In our first analysis, we focus only on Florida students in the third grade in 2001–02 or 2002–03 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.[6] 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 score but who entered the third grade in the year prior to the policy (who were thus were not subjected to the program). That is, our treatment group consists of the first cohort of low-achieving students subject to the test-based retention policy, and our control group consists of similarly low-achieving students who were not subject to the policy because they happened to be born a year earlier. On average, the two groups should be very similar, and any observed differences can be controlled statistically.

[6] The cutoff was an FCAT reading score of 1045 DSS points. DSS points are discussed later in this paper.

We compare the test-score gains of students in the first and second years after their initial third-grade year. For each group of students, we measure the test-score gains that they made between the baseline year and two years afterward. Thus, in the evaluation of gains after one year, we compare the gains that the control group made between 2001–02 and 2002–03 with the gains made by the treatment group between 2002–03 and 2003–04. For the analysis of gains in the second year after retention, we compare the gains that the control group made between 2001–02 and 2003–04 with the gains that the treatment group made between 2002–03 and 2004–05.

The test scores of students in our two comparison groups not only differ 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 (2004–05) 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, and thus in the second year after baseline (2003–04), most of them were in the fifth grade.

The existence of Developmental Scale Scores (DSS) allows us to compare student gains on the FCAT reading test regardless of the year and grade in which the test was administered. These scores were developed by the Florida Department of Education as a uniform measure of proficiency across grades and years. For example, a third-grade student who earns a DSS of 1000 on the FCAT reading test in 2002–03 has the same proficiency as a fourth-grade student who earns a DSS of 1000 on the FCAT reading test in 2004–05. Similar scale scores have also been developed for other commercial standardized tests such as the Stanford testing series. Previous research has shown that the FCAT produces results that are very similar to those of the Stanford-9 test (Greene, Winters, and Forster 2004 West and Peterson 2005).[7]

Table 2 reports descriptive statistics on the treatment and control groups and compares them using a one-way ANOVA analysis. The table shows that the two groups of students are, in fact, statistically different on all observed dimensions. The control group of students with low test scores who entered third grade the year before the policy was in place are slightly more likely to be white or Asian (and consequently less likely to be Hispanic or African-American) and have test scores that are below those of the treatment group. However, though each of these differences is statistically significant, most are quite insubstantial. Only whether the individual is white or whether he is Hispanic differs by more than a single percentage point between the groups. These modest differences that do exist can be controlled statistically.

The across-year comparison approach is 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 has experimented with many educational reforms, including vouchers, charter schools, and other forms of test-based accountability. 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

[7] Greene and Winters (2006) also evaluated the effect of the retention policy on the Stanford-9 as a validity check on the FCAT results. This is no longer available as a comparison in Florida, however, because in 2004–05 the state switched to the Stanford-10, the newest edition of the test. This is further complicated by the fact that not all districts immediately switched to the Stanford-10 and instead continued to administer the Stanford-9 that year. However, there is enough previous research indicating that FCAT results correlate strongly with those of the Stanford series that we can have confidence in the FCAT alone.

provided in the third grade so that fewer students would be retained. The statistically higher baseline reading scores for our treatment group reported in Table 2 indicate that this bias could exist. The difference in baseline test scores highlights the importance of controlling for these scores in all the analyses.

Regression Discontinuity Comparison

For a check on robustness of the results of our across-year approach and to compare our results more directly with those of Roderick and Nagaoka (2005), we further analyze the effect of Florida's retention policy using a regression discontinuity design. The use of regression discontinuity has been growing in popularity as a design for evaluating public policy. This design is useful in cases such as this, when a treatment is primarily determined by the reaching of a threshold of some kind. Van der Klaauw (2001) shows that if obtaining a treatment is conditioned on meeting a certain known threshold, an analysis of individuals in a narrow margin around the threshold approximates random assignment. That is, chance has a large influence over whether students are just above or just below the promotion threshold, so students on either side of the threshold should be very similar at baseline. Differences in their progress over time can then be attributed to whether they happened to be promoted or retained, since the two groups were nearly identical at the start.

We take advantage of the existence of a known cutoff score below which students were more likely to be retained and above which they were more likely to be promoted. The design we utilize is very similar to that used by Roderick and Nagaoka (2005) in their evaluation of Chicago's objective retention policy as well as to other studies outside of education (see, for example, Van der Klaauw 2001, Angrist and Lavy 1999, DiNardo and Lee 2004).

In this evaluation, we compare the test-score gains of students whose reading scores in 2002–03 were just below the threshold required for promotion with students who were in the third grade that same year and whose scores were just above this threshold. Unlike the "across-year" analysis, all students in this design were in the third grade in 2002–03 and were subject to the policy if they did not score above the necessary threshold. Since all students are in the same grade and age cohort, they were all uniformly affected by policies other than the retention policy. Thus, the regression discontinuity approach does not suffer from the limitation of the previous across-year analysis that other policies could affect the results.

In their evaluation of Chicago's policy, Roderick and Nagaoka (2005) use grade-equivalency scores and draw the discontinuity line at scores that were within three months of the threshold.[8] However, DSS scores are not directly convertible into grade equivalents, so we are left to produce our own definition of those "just" above and below the threshold.

Lacking a formal definition for those who are "just" below or above a threshold, we use two potential definition strategies in the regression discontinuity design. We draw the discontinuity first for those whose score on the third-grade FCAT reading test in 2002–03 (the test used for the retention decision) was within 50 DSS points of the threshold for retention and then for whether it was within 25 points of the threshold. In

[8] Grade equivalency is another type of score that allows for comparisons of proficiency across years and grades to which the test was administered. The score is meant to describe the grade level to which a student's proficiency belongs. For example, a grade-equivalency score of 3.5 means that the student had the same proficiency as the median students in the fifth month of third grade.

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. Thus, both definitions of those "close" to the threshold severely limit the sample, and the 25-point definition is quite strict.

The comparison of descriptive statistics of our treatment and control groups using the regression discontinuity cutoffs are recorded in Table 3. Within the 25-point definition of "close," the observed demographic characteristics of the treatment and control groups are statistically identical, except, of course, for their baseline reading-test score and whether or not they were retained. When we compare those within 50 points of the threshold, there are only minor differences in the percentage of students who are white and African-American and who are ineligible for the free or reduced-price lunch program. Thus, the regression discontinuity helps to confirm the robustness of the findings from the across-year model. In particular, the regression discontinuity approach has the advantage of helping to address concerns about unobserved demographic differences between the treatment and control groups in the across-year analysis.

Our method follows the so-called fuzzy discontinuity design, as do many other such papers. That is, the discontinuity of student baseline test scores is not strict. Many students with test scores below the cutoff score were exempted from the policy. Further, some students who scored above the cutoff were nonetheless retained. Table 3 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 25-point definition, the table shows that 59 percent of students with scores below the test-score cutoff were actually promoted (did not receive the treatment) while 4.5 percent of students whose scores were just above the cutoff were actually retained (did receive the treatment).

When there are a lot of exemptions, we risk running into the same methodological dangers that beset earlier studies of discretion-based retention. If exemptions are granted on a discretionary basis, perhaps retained students will once again be incomparable in key unobserved ways. To address this problem, we use a two-stage model. In a two-stage approach, we essentially identify who would have been retained if exemptions did not distort the pool of retained students. Then we predict the effect of this undistorted retention on academic achievement. This technique removes bias that could be introduced by the subjective use of exemptions.

One limitation of the discontinuity approach is that by including only those students whose baseline reading score falls within a very narrow range, we eliminate many potentially useful observations. While our number of observations in the across-year comparison is 78,039 in the second year, under the regression discontinuity this falls to 13,841 under the 50-point threshold and only 7,326 under the 25-point definition.

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 not identical for all students below the retention cutoff--for example, if students with very low baseline proficiency are more or less affected by the policy--then our estimates will not indicate the true effect of retention.

Of course, the across-year design has its limitations as well, such as the danger that different cohorts differ in unobserved ways or are differentially affected by changes in school practices over time. The point of using multiple designs and multiple analyses is to gauge one's confidence in results by seeing if they are robust across different specifications.

Results

The results using multiple research strategies are consistent with the theory that test-based retention of low-proficiency students increases their reading proficiency and that these gains increase over time.

The results of our analyses on the test-score gains made in reading are reported in Table 4. The first column of the table shows the test-score gains in the first year after retention, and the second column shows the test-score gains two years after retention.[9] These results can be interpreted as the gains made by retained students above those made by comparable students who were promoted. Table 4 also contains the results from the three different analyses we performed: the across-year comparison; the discontinuity comparison, using 50 DSS points as the definition of "close" to the promotion threshold; and another discontinuity comparison, using 25 DSS points as the definition of "close."

In both the first and second year, the effects of being retained are statistically significant and positive in all three comparisons. Test-based retention has significant benefits that grow over time and are robust across multiple analytical strategies. In the across-year comparison, the effect of retention on reading scores after one year is small but statistically significant (4.1 DSS points). Two years after students are retained,

[9] The complete models and results are available from the authors upon request.

however, their reading achievement outstrips their counterparts who were promoted by 40.9 DSS points.

These results are confirmed by the regression discontinuity comparisons. In the discontinuity comparison of students whose FCAT reading score was within 50 points of the cutoff score, retained students made test-score improvements over promoted students of 16.3 DSS points in the first year after retention and 57.8 in the second year after retention. We find similar results using the very strict discontinuity comparison of those within 25 points of the promotion threshold. After one year, retained students made reading gains on the FCAT that were 17.9 DSS points higher than students with similar characteristics who were promoted, and these relative gains grew to 60.3DSS points in the second year after retention.

The true size of the retention effect is difficult to interpret from the above results because it is substantially different depending upon the comparison group utilized. This is, however, somewhat to be expected given that the regression discontinuity approach is limited to evaluating only the impact of the policy on those with test scores in a very narrow margin near the cutoff, while the across-year approach measures the impact of the policy for all students who were subjected to it. Thus, the true size of the effect is most likely found in the across-year comparison. However, the fact that in all analyses the effect of retention is positive, highly statistically significant, and grows from the first year to the second year after retention provides confidence that the overall effect of the policy is distinctly positive.

It is also difficult for most people to interpret how large a benefit these improvements in DSS scores really represent. To put them in better perspective, we have converted the results into standard deviations and percentiles in Table 5. A standard deviation is a measure that helps education researchers compare results across different studies that use different tests. A standard deviation represents a portion of a bell curve (or normal curve). If all students were arrayed in a bell curve, 95 percent of them would be within two standard deviations of the average students and 68 percent would be within one standard deviation (more students are packed into the middle of a bell curve).

After one year, retained students benefit by between .01 and .05 standard deviations, depending upon the analysis. These represent small, but statistically significant, effects. After two years, the benefit of retention grows to between .11 and .16 standard deviations, which education researchers would generally regard as moderate benefits. Gains of this size are somewhat smaller than have been observed in evaluations of class-size reduction or voucher programs, which are around one-quarter of a standard deviation, but they are larger than the effects of charter-school programs or increased per-pupil spending, which tend to be between zero and one-tenth of a standard deviation.

While measuring effects in standard deviations permits comparisons with other studies of other programs, these units are still relatively unfamiliar to most non-researchers. To help people understand the magnitude of the effects, we have also converted them into percentiles in Table 5. Percentiles rank all students so that 1 percent would be in each percentile. A student performing at the 50th percentile outperforms 50 percent of all students. Students in our across-year treatment group (those who entered the third grade in 2002–03 with FCAT reading scores below the necessary threshold) had an average score at the 23rd percentile on a nationally normed test also administered to all students in the state. A student at the 23rd percentile outperforms 23 percent of all students but trails the other 77 percent. A gain of five percentile points is easier closer to the middle of the pack, where most students are grouped, and harder on the tails, just as passing other students in a foot race is easier if one is running in the middle of the pack than if one is way ahead or way behind, where there is more distance between each runner. Given that retained students start at the 23rd percentile in reading, they would barely gain one percentile point one year after being retained but would gain between three and 5.1 percentile points two years after being retained.

Comparing Florida with Chicago

Using several analytical strategies, we find that Florida's test-based retention policy has led to significant improvements in reading scores for those students who were retained. These results contradict those of Roderick and Nagaoka (2005), who also found initial benefits after the first year of the program but found that these benefits disappeared in the second year after retention. Because we use a similar basic analytical model as Roderick and Nagaoka, the different results most likely stem from differences in the policies and their implementation in Chicago and Florida, not from differences in the research designs.[10] Although we are unable to test the effects of the different characteristics of the two programs empirically, some key policy differences deserve discussion.

One important difference in the two policies is that Florida's policy regulated and guided the exemptions from the policy while Chicago's policy had no formal rules for

[10] Unlike Roderick and Nagaoka (2005), we do not perform a one-stage HLM design because the one-stage design does not accurately account for the large number of exemptions in the policy's implementation. But we both perform the same type of discontinuity comparison and yet arrive at different conclusions for the different programs.

promotion of students with scores below the minimal threshold. The idea of allowing exemptions in Florida is to accommodate the needs of students whose test scores, for some reason, do not truly demonstrate their academic proficiency or who have some exceptional characteristic that could explain low test scores (such as a disability or limited proficiency in English). If these exemptions effectively promote students for whom retention would be harmful, they would add to the effectiveness of the policy overall. Thus, part of the negative findings in Chicago could be attributed to the fact that the policy in that city retained some students who would have benefited from promotion. Without formal rules for promoting students, it is likely that the exemption strategy was not well tailored to identifying individuals who would benefit from promotion, and it could have been quite arbitrary. In Florida, on the other hand, the procedures for exempting students from retention may have more effectively guided educators about who would benefit most from being exempted from test-based retention.

Another difference between the policies in Chicago and Florida is that the Chicago policy underwent several changes in its implementation, while Florida's policy has remained consistent. Changes in the policy might cause uncertainty in the response of schools and thus inconsistent results. If educators believe that a retained student will be promoted because of a change in the retention policy rather than because of improved skills, their incentives to improve student skills are undermined.

In addition, recent allegations of testing impropriety in Chicago (see Jacob and Levitt 2003and Greene, Winters, and Forster 2004) compared with validation of testing integrity in Florida (see Greene, Winters, and Forster 2004; West and Peterson 2005) may help explain the 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. If that explains the different findings, the lesson would be that test-based promotion with a valid testing system is beneficial while the same policy without testing integrity may be harmful.

Of course, these possible explanations for the differences in the findings in Florida and Chicago are only hypotheses and require further empirical examination. What is clear, however, is that there are differences in the effect of test-based retention across these two jurisdictions and that these differences do not appear to have been caused by variation in the way the programs were evaluated.

Conclusion

While we can have confidence that test-based retention in Florida has academic benefits, we do not know a number of things. We do not know whether the gains we have observed two years after students are retained will continue to hold, expand, or disappear over time. We intend to continue tracking their progress to find out.

We do not know whether test-based retention policies in other school systems, such as Texas and New York City, have benefits similar to those in Florida. The results from Florida tell us that test-based retention when implemented under the right conditions improves student learning, but the evidence from Chicago reminds us that the same policy improperly implemented can be counterproductive. These programs in other school systems need to be carefully evaluated to determine if they are producing benefits or if their features need to be modified to achieve results similar to those found in Florida.

We do not know whether the benefits of test-based retention in Florida justify the additional costs involved. Retaining students means that students may spend an additional year in public schools. With national per-pupil spending topping $10,000, adding another year of school for a large number of students requires significant additional spending over time. Of course, additional spending that significantly improves outcomes for students may well be worth it. Without tracking the benefits over the long term and without a careful cost-benefit analysis, it is difficult to draw conclusions on this.

What we can know is that test-based retention in Florida is helping students improve their reading. This evaluation supports the theory that students with low test scores who are promoted appear to lack the minimum skills to prosper in the next grade. Retaining low-scoring students gives those students a chance to catch up on their skills so that they have the wherewithal to progress academically.

Given the frustrating stagnation in student achievement over the last three decades, despite the significant increase in resources and efforts to improve learning, any large-scale policy that produces progress is promising. Test-based retention should continue to be tried and carefully evaluated to see if this promise can become a reality of higher student achievement for students nationwide.

References

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

DiNardo, J., & Lee, D. S. (2004). "Economic impacts of new unionization on private sector employers: 1984–2001." *Quarterly Journal of Economics*, *119*(4), 1383–1441.

Greene, J. P., & Winters, M. A. (2006). "Getting ahead by staying behind." *Education Next*, *6*(2), 65–70.

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

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

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

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

Jimerson, S. R. (2001). "Meta-analysis of grade retention research: Implications for practice in the 21st century." *School Psychology Review*, *30*(3), 420–37.

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

Van der Klaauw, W. (2001). "Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach." *International Economic Review*, *43*(4), 1249–87.

West, M. R., & Peterson, P. E. (2005). "The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments." Paper presented before the annual conference of the Royal Economic Society, University of Nottingham, March 23, 2005.

Table 1		
Promotion Characteristics -- All Students in Third Grade in 2002-03 with Scores Below Test-Score Threshold		
Exemption for:		
	Percent	
Promoted Because…		
No Code Listed	4%	
Limited English Proficient	6%	
Disability -- Testing Not Appropriate	0%	
Passed Alternative Test	7%	
Student Portfolio	3%	
Disablity -- Has Received Extensive Instruction	7%	
Already Retained Twice	1%	
No Longer Enrolled in School System	3%	
No Explanation	13%	
Total Promoted	43%	
Retained	57%	
* Totals may not sum due to rounding		

Table 2			
Comparison of Descriptive Statistics -- Across-Year Comparison			
	Control - 3rd grade 2001-02, reading score below threshold	Treatment - 3rd grade 2002-03, reading score below threshold	
Indian	0.2%	0.2%	*
Asian	1.1%	1.0%	*
African-American	37.5%	36.6%	*
Hispanic	27.9%	30.0%	*
Multiple Race	1.8%	2.0%	*
White	31.4%	30.2%	*
Ineligible for Free or Reduced-Price Lunch	73.1%	76.0%	*
Limited English Proficient	26.1%	26.6%	*
Baseline DSS Reading Score	761	776	*
Retained in Baseline Year	6.3%	56.8%	*
N	47,684	40,881	
* indicates statistically different at .05 level			

Table 3

Comparison of Descriptive Statistics -- Regression Discontinuity Analysis

	Within 50 Points Above or Below Threshold				Within 25 Points Above or Below Threshold			
	Above	Below			Above	Below		
Baseline Reading Score	1073	1022	*		1060	1033	*	
Proficient in English	75.8%	74.8%			74.8%	75.3%		
Asian	1.2%	1.2%			1.3%	1.2%		
African-American	34.0%	35.4%			35.8%	35.7%		
Hispanic	25.5%	26.4%			25.5%	25.9%		
Indian	0.3%	0.3%			0.3%	0.4%		
Multiple Race	2.5%	2.3%			2.2%	2.3%		
White	36.2%	34.2%	*		34.6%	34.5%		
Ineligible for Free or Reduced-Price Lunch	24.5%	22.2%	*		23.2%	23.1%		
Retained in Baseline Year	4.2%	43.3%	*		4.5%	41.4%	*	
N	7,871	7,362			3,826	4,267		

* indicates statistically different at .05 level

Table 4					
Effect of Retention on Reading Developmental Scale Scores					
	1-Year Gain		2-Year Gain		
Across-Year Comparison	4.1		40.9		
N	79,747		78,039		
Adjusted R-Square	0.17		0.24		
Regression Discontinuity -- Within 50 Points	16.3		57.8		
N	14,172		13,841		
Adjusted R-Square	0.03		0.08		
Regression Discontinuity -- Within 25 Points	17.9		60.3		
N	7,501		7,326		
Adjusted R-Square	0.03		0.08		
Controlling for race, free lunch status, limited English proficiency, baseline test scores, and school district dummy					

Table 5

Effect of Retention on Reading in Standard Deviation and Percentiles*

	Standard Deviation		Percentiles*	
	1-Year Gain	2-Year Gain	1-Year Gain	2-Year Gain
Across-Year Comparison	0.01	0.11	0.3	3.4
Regression Discontinuity -- Within 50 Points	0.04	0.15	1.2	4.8
Regression Discontinuity -- Within 25 Points	0.05	0.16	1.5	5.1

*Assuming student began at the 23rd percentile