Education Working Paper

No. 7 December 2004

An Evaluation of Florida's Program to End Social Promotion

Jay P. Greene, Ph.D.

Senior Fellow, Manhattan Institute for Policy Research

Marcus A. Winters

Research Associate, Manhattan Institute for Policy Research

CENTER FOR CIVIC INNOVATION

AT THE MANHATTAN INSTITUTE

EXECUTIVE SUMMARY

Nine states and three of the nation's biggest cities have adopted mandates intended to end "social promotion"—promoting students to the next grade level regardless of their academic proficiency. These policies require students in certain grades to reach a minimum benchmark on a standardized test in order to move on to the next grade. Florida, Texas, and seven other states, as well as the cities of New York, Chicago, and Philadelphia, have adopted mandatory promotion tests; these school systems encompass 30% of all U.S. public-school students. Proponents of such policies claim that students must possess basic skills in order to succeed in higher grades, while opponents argue that holding students back discourages them and only pushes them further behind.

This study uses individual-level data provided by the Florida Department of Education to evaluate the initial effects of Florida's policy requiring students to reach a minimum threshold on the reading portion of the Florida Comprehensive Assessment Test (FCAT) to be promoted to the 4th grade. It examines the gains made in one year on math and reading tests by all Florida 3rd graders in the first cohort subject to the retention policy who scored below the necessary threshold, comparing them to all Florida 3rd graders in the previous year with the same low test scores, for whom the policy was not yet in force. Because some students subject to the policy obtained special exemptions and were promoted, the study also uses an instrumental regression analysis to separately measure the effects of actually being retained. The study measures gains made by students on both the high-stakes FCAT and the Stanford-9, a nationally respected standardized test that is also administered to all Florida students, but with no stakes tied to the results.

The authors intend to follow the same two cohorts of students in future studies to evaluate the effects of this new policy over time. The findings of this study, evaluating Florida's program after its first year, include:

- Low-performing students subject to the retention policy made gains in reading greater than those of similar students not subject to the policy by 1.85 percentile points on both the FCAT and the Stanford-9.
- Low-performing students subject to the retention policy made gains in math greater than those of similar students not subject to the policy by 4.76 percentile points on the FCAT and 4.43 percentile points on the Stanford-9.
- Low-performing students who were actually retained made gains in reading greater than those of similar students who were promoted by 4.10 percentile points on the FCAT and 3.45 percentile points on the Stanford-9.
- Low-performing students who were retained made gains in math greater than those of similar students who were promoted by 9.98 percentile points on the FCAT and 9.26 percentile points on the Stanford-9.

About the Authors

Jay P. Greene is a Senior Fellow at the Manhattan Institute's Education Research Office, where he conducts research and writes about education policy. He has conducted evaluations of school choice and accountability programs in Florida, Charlotte, Milwaukee, Cleveland, and San Antonio. He has also recently published research on high school graduation rates, charter schools, and special education.

His research was cited four times in the Supreme Court's opinions in the landmark *Zelman v. Simmons-Harris* case on school vouchers. His articles have appeared in policy journals, such as *The Public Interest*, *City Journal*, and *Education Next*, in academic journals, such as *The Georgetown Public Policy Review*, *Education and Urban Society*, and *The British Journal of Political Science*, as well as in major newspapers, such as the *Wall Street Journal* and the *Washington Post*.

Greene has been a professor of government at the University of Texas at Austin and the University of Houston. He received his B.A. in history from Tufts University in 1988 and his Ph.D. from the Government Department at Harvard University in 1995. He lives with his wife and three children in Weston, Florida.

Marcus A. Winters is a research associate at the Manhattan Institute's Education Research Office, where he studies and writes on education policy. He has co-authored several studies on a variety of education policy issues including high-stakes testing, charter schools, and the effects of vouchers on the public school system. His op-ed articles have appeared in numerous newspapers, including the *Washington Post*, the *Christian Science Monitor*, and the *San Francisco Chronicle*. He received his B.A. in political science with departmental honors from Ohio University in 2002.

Acknowledgments

The authors would like to thank the Florida Department of Education, especially the staff and administration of its K-20 Data Warehouse, for compiling and making available the data necessary for this study.

About Education Working Papers

A working paper is a common way for academic researchers to make the results of their studies available to others as early as possible. This allows other academics and the public to benefit from having the research available without unnecessary delay. Working papers are often submitted to peer-reviewed academic journals for later publication.

Table of Contents

An Evaluation of Florida's Program to End Social Promotion

INTRODUCTION

School systems across the nation have recently enacted substantial new programs to stop schools from promoting students from grade to grade regardless of academic proficiency. To end this practice, known as "social promotion," several large school systems now require students in particular grades to demonstrate a benchmark level of mastery in basic skills by passing a standardized test before they can be promoted. These controversial mandates have been adopted by nine states, including Florida and Texas, as well as by New York City, Chicago, Philadelphia, and other cities.

Proponents of these programs think that schools do students no favor by promoting them to the next grade if they do not possess the skills necessary to succeed at a higher level. They argue that if a student lacks basic proficiency in reading concepts at the third-grade level, that student will certainly fail to grasp concepts intended for fourth-graders. On this view, once a student is promoted beyond his skills he will only continue to fall further behind as material becomes more difficult in later years.

While these arguments may be plausible, there is currently no research backing them up. Those opposed to ending social promotion, on the other hand, point to a wide body of research suggesting that students who are retained in a grade for an extra year are academically and emotionally harmed by the experience. Several studies have indicated that students who are held back have lower test scores and are more likely to drop out than similar counterparts who are not held back.

However, prior research on grade retention is severely limited by methodological problems that are unavoidable in evaluating retention policies based on subjective criteria (i.e., teachers' evaluations that students should be retained). Furthermore, it is questionable whether research on students who were retained according to subjective criteria is even relevant in the first place to retention policies based on objective criteria. For example, it is possible that the potentially harmful stigma currently associated with retention might not apply to the same extent under the new system, which holds back much larger numbers of students. It is certainly possible that retaining thousands of students according to their standardized test scores might influence student outcomes in ways far different from previous retention practices that singled out a very small number of students for retention based upon subjective criteria. New research looking directly at the effectiveness of test-score mandates intended to end social promotion is necessary in order for policymakers and the public to make informed decisions.

This study seeks to provide research to inform that debate by evaluating Florida's early experience with ending social promotion through standardized testing. Under a law passed by the state legislature, third-graders in Florida must score at the Level 2 benchmark or above on the reading portion of the state's high-stakes test, the Florida Comprehensive Assessment Test (FCAT), in order to be promoted to the fourth grade. Students who fail to reach this benchmark are given supplemental instruction and, unless they acquire an exemption, must repeat the third grade. The third-grade class of 2002–03 was the first affected by the mandate.

Our data were the individual test scores of all third-grade students who failed to reach the minimum benchmark on the FCAT reading test during the 2001–02 and 2002–03 school years. We examined the test-score gains made by students over one year after the point when they failed to reach the benchmark. We included third-graders who missed the benchmark in 2002–03, the year in which the new policy first took effect, as well as third-graders who missed the benchmark in the previous year, when the policy was not yet in effect.

We performed two analyses. Our first analysis measures the effect that being subject to the new program has on student achievement. It compares low-scoring third-graders in 2002–03, who were subject to the program, with low-scoring third-graders from the previous year, who were not. However, some students who were subject to the program received special exemptions that allowed them to advance to the fourth grade in spite of their test scores. Thus, not all students who were subject to the new program were actually retained. So we performed a second analysis in order to measure the effect of actual retention—whether under the new program or under the old retention policy—on low-performing students. We used an instrumental analysis method to compare students in both years who were actually retained with students in both years who were not actually retained.

We find that the early stage of Florida's policy to end social promotion has improved academic proficiency. Our first analysis finds that low-performing students subject to the program made modest improvements in reading and substantial improvements in math compared with those made by low-performing students in the previous year's cohort who were not subject to the program because it had not yet taken effect. Our second analysis finds that the effect of actually being retained is even stronger. We find that low-performing students who were actually retained make relatively large improvements in reading and exceptional improvements in math compared with similarly low-performing students who were promoted.

The findings of this study are encouraging for the use of standardized testing policies to end social promotion, but they are also limited because we are only able to evaluate the effects of the first year of the program. It is certainly possible that the gains made by students affected by the program might not hold up later in their academic careers, as proponents of the policy expect. On the other hand, it is also possible that the gap between students who were socially promoted and those who were retained might widen further as they enter higher grades and the material becomes even more challenging. Further research following these same groups of students will be necessary to track the effectiveness of Florida's retention program over time. For the time being, this study indicates that the use of objective testing to end social promotion leads to substantial academic gains for low-performing students, though we cannot yet determine how long these gains will persist.

That limitation notwithstanding, this study has important implications not only in Florida but nationwide. Since Florida's program is very similar to programs in eight other states, as well as New York, Chicago, and Philadelphia, that also focus on achievement on standardized reading tests, the results of our Florida analysis are likely to indicate the effects of the same policy in these other school systems. Much could be learned by directly measuring those other programs using the method employed in this study, but until such research has been performed, this study provides valuable information that can reasonably be applied to other existing programs and possible future programs as well.

PREVIOUS RESEARCH

Over the last several decades, many studies have examined the effect of retention on future student achievement. Most of these studies have found that student outcomes are negatively affected by retention. However, the quality of these studies is far lower than their quantity.

Holmes (1989) performed an often-cited meta-analysis of the research on grade retention. A meta-analysis is a study of studies: it analyzes the results of multiple previous studies on a particular research question—in this case, the effect of retaining students on future outcomes. The purpose of a meta-analysis is to empirically produce a single cumulative finding from a wide body of research.

In his meta-analysis, Holmes included 63 studies on grade retention with a total of 861 findings. Of the 63 studies he evaluated, 54 reported overall negative effects on grade retention. Of the studies that directly measured academic achievement, Holmes determined that their cumulative finding was that retained students performed 0.19 standard deviation units below promoted students. He concluded that "the weight of empirical evidence argues against grade retention" (p. 28).

While Holmes's meta-analysis is often treated as definitive, some researchers have pointed out serious flaws with his finding (Reynolds 1992 and Alexander,

Entwisle, and Dauber 2003). They argue that the studies in Holmes's meta-analysis are not of high enough quality to support definitive conclusions.

The most serious limitation of previous research on retention is the lack of an adequate control group that can be compared with retained students. Reynolds, whose own study concludes that retention is harmful, points out that "only 25 of the 63 retention studies [included in Holmes's meta-analysis] used matched control group designs (matched prior to data analysis or statistically controlled). Only 16 studies matched students on prior achievement, and only 4 studies matched students on attributes that are consistently found to be predictive of the decision to retain" (p. 102).

So 38 of Holmes's 63 studies did not even have a control group with whose performance retained students were compared. And the other 25 studies, while they do compare retained students with other students, are drawing comparisons with students who are not adequately comparable to retained students.

The problem of finding an adequate control group that can be compared with retained students has not been easy to solve in previous studies. Some past researchers have made great efforts to develop adequate comparison groups, but these efforts have been rendered futile by the subjectivity of grade-retention decisions. In the past, the retention of a student has largely been the result of a teacher's subjective assessment of his ability to succeed at the next level. Therefore, we can expect that students who were retained are fundamentally different from students who were promoted, even if they are similar in all measurable factors such as race, because their teachers evaluated them as being fundamentally different. Further complicating matters is that assessments of students are likely to differ greatly not only among teachers but also among a single teacher's evaluations of various children. In previous studies, the students who were retained are simply not comparable with the promoted students with whom they are compared.

The existence of an objective retention policy in Florida allows for the development of an adequate comparison group not available in previous evaluations. Unlike previous studies, this study compares students who were subject to retention with other students who we know would have been subject to retention had they only been born a year later.

Reynolds also points out that most of the previous studies included in Holmes's meta-analysis have evaluated the effect of retention on white middle-class students in suburban or rural schools (p. 103). Such studies might tell us very little about the effects of retention on urban minority students, whom the new retention policies are most often targeted toward helping and who are, in fact, the most likely to be retained under them.

Previous research supporting retention policies has also suffered from methodological flaws. Many of the studies in Holmes's meta-analysis that find positive effects from retention make within-grade comparisons of students instead of within-age comparisons (see also Alexander, Entwisle, and Dauber 2003). Within-grade studies compare retained students with other students in their cohort grade after they are held back, while within-age evaluations compare retained students with students in their original class who were promoted to the next grade.

Alexander, Entwisle, and Dauber argue that within-grade comparisons are preferable because "comparing repeaters with children who have been exposed to a more advanced curriculum puts them at a decided disadvantage, and a same-age frame of reference almost preordains results that favor promotion" (p. 19). However, the only meaningful way to evaluate the effects of retention is to compare the academic achievement of retained students with an estimate of what their performance would have been had they not been retained. Only a within-age comparison can provide such an evaluation. Within-grade comparisons, even those that evaluate students several years after retention, fail to provide any information about the level at which retained students would have performed had they been promoted.

Even if all this previous research were of high quality, there are strong theoretical reasons to believe that these previous studies, which examine the effectiveness of retention based on subjective criteria, might have little relevance to programs intended to end social promotion by applying objective criteria for retention. If, as several of the researchers who have found retention to be harmful have hypothesized, a

retained student performs worse because he feels excluded and thus inferior, then a policy that holds back thousands of students might dilute this sense of being singled out, limiting the psychological harm associated with retention. Also, subjective assessments of students are vulnerable to inappropriate influences, including teachers' prejudices and pressure brought by parents, in ways that objective criteria of performance might limit. Implementing objective standards, even if they are accompanied by subjective exemptions from those standards for some students, might significantly change the effects of retention in ways that previous research cannot anticipate.

Scholars at the Consortium on Chicago School Research have performed a series of evaluations of that city's objective program to end social promotion through testing (Nagoaka and Roderick 2004). Though Chicago's policy now includes several ways that a student can be promoted with low test scores, the cohorts of students examined by Nagoaka and Roderick in the third, sixth, and eighth grades were required to exceed benchmarks on the Iowa Test of Basic Skills (ITBS), a nationally respected and widely administered standardized test, in order to be promoted to the next grade. In the latest of these studies, conducted in 2004, Nagoaka and Roderick compared the performance of third- and sixth-grade students who scored just below the benchmark on the ITBS, most of whom were retained because of the mandate, with the performance of students who scored just above the benchmark, most of whom were promoted. Nagoaka and Roderick conducted two analyses, similar to the two analyses in our study. First, they measured improvements in performance between the two groups without accounting for whether each student was actually retained, and then statistically adjusted for whether each student was retained or promoted. They were able to measure test-score performance for two years after the implementation of the program.

Nagoaka and Roderick's study found that, after two years of the retention policy, third-grade students were not affected and sixth-grade students were negatively affected by the policy in their performance on the ITBS reading test. However, while their study provides valuable evidence on the effectiveness of Chicago's retention program, it is limited by several factors.

The most important limitation of the Nagoaka and Roderick study is that comparing students in the same cohort who score just above and just below the benchmark is an inadequate method. First, we know that the two groups of students are systematically different from each other because they performed differently on the ITBS. This incomparability could fundamentally distort their results. Our study compares all third-grade students affected by the first year of Florida's social promotion mandate with all those in the prior year who would have been subjected to the mandate had it been in force at that time.

Nagoaka and Roderick argue that doing such a comparison in Chicago would have been inappropriate because standardized test scores were rising throughout Chicago from year to year, especially in the grades that they were evaluating. Because fewer students received failing grades in the cohort that was subject to the new policy, they argue, their treatment group would have been biased. It would have been made up of students who still could not pass the benchmark despite the general improvements in Chicago. Unfortunately, the method that they do use incorporates a worse bias. Students from a single year prior to the implementation of the policy would likely be far more comparable with their treatment group than students in the same year who scored above the benchmark that their treatment group failed to reach.

Comparing similarly low-performing students in different cohort years is also preferable because it allows an evaluation of all students affected by the program. Because their study compares only those students just above and just below the ITBS benchmark, Nagoaka and Roderick's evaluation tells us nothing about the policy's effect on students whose scores were far below the necessary benchmark. Such very low-performing students, who made up about 50 percent of their cohort of sixth-graders in 2000, were simply excluded from their analysis.

The Chicago study is also limited because it only evaluated performance on the ITBS reading assessment. Nagoaka and Roderick argue that they used only reading because the vast majority of students were retained because of their reading scores, not their math scores. But it is quite possible, perhaps even probable, that students who were retained because of their reading scores might make larger gains

in math than in reading compared with students who were promoted without the necessary skills. This might seem counterintuitive given that such programs are usually portrayed as being targeted to improving literacy, not numeracy. In fact, similar programs elsewhere, such as in Florida, only require students to pass the reading assessment to be promoted. Nonetheless, retained students might make greater academic progress in math because learning in that subject is more cumulative than reading is. For example, if a student does not adequately learn addition, he will be particularly unlikely to adequately learn multiplication, because understanding the latter requires mastery of the former. While learning in reading may have a similar cumulative effect, it is likely not to be as dramatic as for math.

FLORIDA'S PROGRAM TO END SOCIAL PROMOTION

Over the last several years, Florida has attempted to make substantial reforms to its struggling school system, which consistently ranks close to the bottom on nearly all academic indicators.[1] In May 2002, the legislature decided to focus its attention on the problem of social promotion—the practice of promoting students to the next grade level independent of their academic proficiency—at the end of the third-grade year. Those opposed to social promotion argued that students who leave third grade without reaching a certain minimum benchmark of basic skills will fail to adequately grasp the more difficult curriculum of fourth grade. They further argued that the gap between students with and without basic skills would continue to grow as material continued to become progressively harder over time because socially promoted students would lack the foundation on which to build their body of knowledge. They claimed that students who cannot read at a proficient level at the end of third grade would benefit in both the short and long run by retaking the same material again instead of moving to a higher grade with more difficult material.

Florida revised its school code to require third-grade students to score at the Level 2 benchmark or above on the reading portion of the FCAT, which was already used throughout the state as a high-stakes standardized test, in order to be promoted to the fourth grade. By requiring that all students possessed at least the basic proficiency necessary to succeed in the next grade level, the reformers hoped that ending social promotion would lead to great academic gains. The third-grade class of 2002–03 was the first to be affected by the law.

The law allowed for some exceptions to the retention policy. A child who misses the FCAT benchmark can be exempted from the policy and promoted to fourth grade if he meets any one of the following criteria: 1) he is a Limited English Proficiency student who has received less than two years of instruction in an English for Speakers of Other Languages program; 2) he has a disability sufficiently severe that it is deemed inappropriate for him to take the test; 3) he demonstrates proficiency on another standardized test; 4) he demonstrates proficiency through a performance portfolio; 5) he has a disability and has received remediation for more than two years; or 6) he has already been held back for two years.[2] Of third-grade students in 2002–03 who scored below the Level 2 threshold and were thus subject to retention under the new policy, 21.3 percent were reported as having received one of these exemptions.[3]

Florida's policy is similar to those recently enacted by other large school systems. As in Florida, all third-grade students in Texas must pass the reading portion of that state's standardized test to be promoted to the fourth grade. Seven other states have adopted similar policies in various grades. New York City has a similar reading mandate for third-graders and has recently expanded its mandate to require fifth-grade students to pass a standardized test to earn promotion as well. Chicago uses whether students pass the reading and math sections of the Iowa Test of Basic Skills in the third, sixth, and eighth grades as a strong component of retention decisions. Philadelphia is now adopting mandatory exams for promotion in grades three through eight. These nine states and three cities enroll a full 30% of all students in public schools in the U.S.[4]

METHOD

Our data include low-scoring students from two school years. First, we include all Florida students who entered the third grade for the first time in 2002–03 and scored below the Level 2 threshold on the FCAT reading test in that year.[5] This was the

first cohort of students in the state subject to the policy requiring them to pass the FCAT reading test in order to be promoted. Our study includes all students who did not pass the FCAT reading test; however, because exemptions from the new policy were available, many of the students we include were not actually retained. Of the third-graders in 2002–03 included in our study for whom we have necessary test scores for our analysis, 60 percent were actually retained. We also include all students who entered third grade for the first time in the 2001–02 school year and who also scored below Level 2 on the FCAT reading test. These students had test scores that would have made them subject to the new policy's retention mandate had it been in effect in that year. Of third-graders in 2001–02 included in our study for whom we have necessary test scores for our analysis, 8.7 percent were retained. The students from both school years are very similar in all respects except for the year in which they happened to have been born, making comparisons between their improvements particularly meaningful.

We analyzed the one-year test-score gains that students made on state-mandated math and reading tests. The existence of developmental-scale scores on each of the tests allows us to compare the test-score gains of all the students in our study even though they took different tests designed for different grade levels. Developmental-scale scores are designed to measure academic proficiency on a single scale for students of any grade and in any year. For example, a third-grader with a developmental-scale score of 1,000 and a fourth-grader with a developmental-scale score of 1,000 have the same level of academic achievement; if a student gets a developmental-scale score of 1,000 in 2001–02 and gets the same score of 1,000 in 2002–03, this indicates that the student has not made any academic progress in the intervening year.

We analyzed the improvements made by students over one year in math and reading scores on the criterion-referenced as well as norm-referenced versions of the FCAT. Both of these are standardized tests that Florida students are required to take. For purposes of clarity, throughout the rest of this study we will follow widespread practice and refer to the criterion-referenced version of the test as the "FCAT" and the norm-referenced version as the "Stanford-9."[6]

Each year, all Florida students in grades three through ten take both the FCAT and the Stanford-9. All students in grades three through ten take both the math and reading sections of both tests each year; other subjects are also tested intermittently. The reading portion of the FCAT is the test that third-grade students must pass in order to be promoted. There are other high stakes tied to the results of the FCAT as well. Every year, the state grades each school from A to F, based primarily on its students' performance on the FCAT. The Stanford-9 is a highly respected standardized test that is frequently administered by states and school districts across the nation. Florida does not attach meaningful stakes to the results of the Stanford-9, as it does to the FCAT. The Stanford-9 is administered to help parents better understand their children's proficiency levels and has been used by researchers and reporters to check the reliability of the results of the high-stakes FCAT (see Greene, Winters, and Forster 2002 and Harrison 2004).

The existence of the Stanford-9 is particularly helpful for our analysis. Several researchers argue that the results of high-stakes tests like the FCAT are routinely distorted because they create adverse incentives for teachers and school systems to manufacture high scores either by "teaching to the test"—changing curriculum and teaching practices in such a way as to raise test scores without increasing real learning (for example, see Amrein and Berliner 2002, Klein et al. 2000, McNeil and Valenzuela 2000, Haney 2000, and Koretz and Barron 1998)—or by outright cheating (for example, see Cizek 2001, Dewan 1999, Hoff 1999, and Lawton 1996). The absence of any substantial consequences attached to the results of the Stanford-9 helps to remove these concerns for our analysis. Since there are no meaningful stakes tied to it, there is no particular incentive for teachers or school systems to attempt to manipulate its results. Thus, if we find similar results on both the high-stakes FCAT and the low-stakes Stanford-9, we can be confident that our findings indicate improvements in real learning and are not distorted by adverse incentives created by high-stakes testing.[7]

With the cooperation of the Florida Department of Education, we obtained individual student-level test scores on the math and reading sections of the FCAT and Stanford-9 for the entire population of students in the state of Florida who met the necessary criteria to be part of our study. We obtained test-score and

demographic information for all students in the state of Florida who first entered third grade in 2001–02 and scored below the Level 2 threshold on the FCAT reading test in that year, as well as for all Florida students who entered the third grade in 2002–03 and scored below Level 2 on the FCAT reading test in that year. The developmental-scale scores required to reach Level 2 on the FCAT reading test were consistent for each year's cohort. For each student in our analysis, we also collected data on race, free or reduced-price school lunch status, and whether the student was considered Limited English Proficient.

We calculated the developmental-scale-score gains on the FCAT and Stanford-9 made in each student's first third-grade year and the following year. For the students affected by the retention policy, we measured the test-score gains they made between the 2002–03 and 2003–04 administrations of the tests. For students who were not affected by the program, we measured their test-score gains between the 2001–02 and 2002–03 administrations of the tests. For students in each group, our calculations of test-score gains were independent of whether the student was administered the third-grade test (indicating that the student was retained) or the fourth-grade test (indicating that the student was promoted). Since developmental-scale scores are consistent between year and grade, the gains we calculated are equivalent for all students.

Our first analysis measured the effect of Florida's retention policy. For this analysis, we were not concerned with whether students who were subject to the retention policy were actually retained or received an exemption and were promoted to the next grade. The availability of exemptions is a meaningful part of the retention policy and thus should be included in its evaluation. The state's policy is intended as a treatment for every third-grade student who scored below the necessary benchmark on the FCAT, even those students who earned an exemption.

To measure the effect of the program, we performed a linear regression comparing the developmental-scale-score gains made by our treatment group, students who first entered third grade in 2002–03 and scored below the FCAT benchmark in that year, with our control group, students who first entered third grade in 2001–02 and scored below the FCAT benchmark in that year. In this regression, we controlled for dummy variables indicating each student's race, whether the student received a free or reduced-price school lunch, and whether the student was deemed Limited English Proficient.[8] We also controlled for each student's test score during his first third-grade year, providing a control for the baseline test performance for each student.

For our second analysis, we evaluated the effect of actually retaining low-performing students. Here we compared low-scoring students from either year who were actually retained with low-scoring students from either year who were not actually retained. We performed a two-stage least squares regression analysis, where the variable of interest was whether a student was retained or promoted. This model uses student demographics and an exogenous variable—the cohort to which a student belongs, which for all intents and purposes is determined by the year in which a student was born—to predict whether each student will be retained. It then uses that prediction to measure the relationship between retention and test-score improvements. We again controlled for student race, free or reduced-price school lunch status, Limited English Proficiency status, and baseline test scores.

In addition to performing these two analyses for the general student population, we also performed each analysis for racial subgroups. This allows us to examine whether the retention policy, or actually retaining students, is having a different effect on students of different races.

RESULTS

Our results are encouraging for the effectiveness of policies that retain students based on standardized test scores. After the first year of the program, we find that the performance of students who were subject to the retention policy, or who were actually retained, exceeded the performance of students who were not subject to the retention policy, or who were not actually retained. Our results are also remarkably consistent between the high-stakes FCAT and the low-stakes Stanford-9, indicating that they have not been tainted by manipulations of the high-stakes testing system.

When interpreting the results of our analysis, it is important to understand that while developmental-scale scores are consistent on a test between grades

and years, they are not consistent between subjects (reading and math) or between the two different standardized tests (the FCAT and the Stanford-9). For example, an increase of ten developmental-scale points is far greater on the Stanford-9 than on the FCAT. This has no effect on our analysis, but it should be kept in mind when interpreting our results. To facilitate comparisons between results on the different subjects and tests, in addition to reporting the difference in developmental-scale scores for each test result, we also report in Table 1 the number of standard deviation units that this difference represents. Standard deviation units are equivalent between the subjects and tests and allow for more meaningful comparisons across subjects and between the FCAT and Stanford-9.

Given that some readers may be unfamiliar with standard deviation units, we also convert all results into the equivalent gain in percentile points in Table 2. Standard deviation units represent a certain amount of change in scores among students whose results are distributed along a normal curve. Assuming that students begin at the 23rd percentile, which was the average Stanford-9 reading score for the students included in this study, we can calculate the percentile point equivalent of a change in standard deviation units by measuring how far students would advance along the normal curve distribution.

As the summary of our findings in Tables 1 and 2 show, students subject to the retention policy make gains of 0.06 standard deviation units in reading and between 0.14 and 0.15 standard deviation units in math relative to students not subject to that policy. Those benefits translate into about two percentile points on reading and five percentile points on math over a one-year period for the average student in our study. Students actually retained made gains of 0.11 to 0.13 standard deviation units on reading and 0.28 to 0.30 standard deviation units on math relative to students who were promoted. Those benefits translate into a benefit of about three or four percentile points in reading and about nine or ten percentile points in math over a one-year period for the average student in our study.

Tables 3 and 4 report the results on the FCAT and Stanford-9 reading tests, respectively, in our first analysis, evaluating the effects of the retention program without accounting for whether students were actually retained. Table 3 shows that on the FCAT, students in our treatment group—those affected by the retention policy—made reading-test-score improvements that were 16.66 developmental-scale points greater than those of students in our control group, who were not affected by the policy. This translates into a gain of about 0.06 standard deviation units for our treatment group on the FCAT reading test, which is a modest improvement after a single year. Our results on the Stanford-9 reading test, reported in Table 4, are remarkably similar. Students subject to the retention policy made test-score improvements that were 1.44 developmental-scale points greater than those of our control group on the Stanford-9 reading test, which also translates into a difference of about 0.06 standard deviation units. Both these results are statistically significant at a very high level (p-values < 0.001).

The findings for our first evaluation in math on the FCAT and the Stanford-9, reported in Tables 5 and 6, respectively, are even greater. On the FCAT math test, as reported in Table 5, students subject to the retention policy made improvements that were 41.67 developmental-scale points greater than those of our control group, translating to an increase of about 0.15 standard deviation units. We again find very similar results on the Stanford-9. Table 6 shows that on the Stanford-9 math test, students subject to the policy outperformed our control group by an average of 4.50 developmental-scale points, also translating to about 0.14 standard deviation units. Each of these results is also statistically significant at a very high level (p-values < 0.001).

We next performed our evaluation of the effect of actually retaining students. Our results for the effect of retention on reading scores are reported in Tables 7 and 8. On the FCAT reading test, as indicated in Table 7, students who were retained made improvements that were 32.48 developmental-scale points, or 0.13 standard deviation units, greater than those made by students who were promoted. Table 8 shows that on the Stanford-9 reading test, students who were retained improved by 2.80 developmental-scale points more than students who were promoted, which translates to a gain of about 0.11 standard deviation units. Both these findings are statistically significant at a very high level (p-values < 0.001).

Tables 9 and 10 show that the results of our analyses comparing retained students with promoted students are larger in math. Table 9 shows that retained students improved by 82.54 developmental-scale points more than promoted students on the FCAT math test, an improvement of about 0.30 standard deviation units. On the Stanford-9 math test, Table 10 indicates that retained students improved by an average of 8.77 developmental-scale points more than promoted students, which translates to a difference of about 0.28 standard deviation units. Again, both results are statistically significant at a very high level (p-values < 0.001).

We also performed both types of analysis on all racial subgroups in the student population. In both cases, we found very similar results for students in each racial group. Both the retention policy and actual retention of students have positive effects of about the same magnitude on students of all races.

DISCUSSION

The results of our analyses are encouraging for the use of objective retention policies based on the results of standardized tests. Each of our analyses found consistently positive results for the use of such retention policies.

The gains made both by students subject to retention and those who were actually held back are substantial compared with gains that we would expect from other popular education reforms. For example, the Tennessee Star Project's widely cited study of class-size reduction found that reducing class sizes from about twenty-four students per teacher to about fifteen students per teacher led to a statistically significant increase of about 0.2 standard deviation units.[9] The highest-quality research on school vouchers has found that the use of such scholarships to attend private schools leads to a gain of two to eleven percentile points.[10] We find that subjecting students to a retention policy improves scores by about 0.06 standard deviation units (about 1.85 percentile points) in reading and about 0.15 standard deviation units (about 4.76 percentile points) in math after one year. We also find that actually retaining students leads to a gain of about 0.12 standard deviation units in reading (about four percentile points) and a gain of about 0.3 standard deviation units in math (about nine percentile points) after one year. This indicates that Florida's retention policy is likely at least as effective at increasing student test scores as these other popular reforms.

That our results are consistently uniform between the FCAT and the Stanford-9 indicates that they are not distorted by perverse incentives for schools and teachers to manipulate their test results. Many argue that teachers and schools will respond to the new retention policy by manipulating test scores, either directly by cheating or indirectly by teaching students skills that will help them to improve their test scores but will not provide real academic proficiency. This argument would only have merit if we found strong gains on the high-stakes FCAT and no similar gains on the low-stakes Stanford-9, on which there is no incentive to manipulate scores. If teachers are in fact changing their curricula with the intent to "teach to" the FCAT, they are doing so in ways that also contribute to gains on the highly respected Stanford-9. This indicates that whatever changes teachers have made have resulted in real increases in students' proficiency.

Some might be surprised that in both analyses, we found much greater gains in math than in reading. This might seem particularly odd, given that it is the reading portion of the FCAT that students must pass to earn promotion and that the rhetoric supporting Florida's retention program emphasizes that it will improve student literacy.

First, it is important to remember that we do find modest improvements in reading (about 0.06 standard deviation units) for students who were subject to the policy, and relatively large improvements (about 0.13 standard deviation units) for students who were actually retained. The larger test-score gains in math do not in any way imply that the program is failing to live up to expectations in reading.

Furthermore, there are strong theoretical reasons to believe that retained students would make greater improvements in math than in reading. It is reasonable to assume that most students whose reading skills are deficient will also be likely to have low proficiency in math. If retention is beneficial for students with low proficiency, we would expect them to make gains in both subjects. Also, the skills required in math may be more cumulative than those

required in reading. For example, a student who cannot add properly is very unlikely to adequately learn multiplication because the latter requires knowledge of the former. Reading instruction is also cumulative, but it is likely to be less so than math. Thus, reviewing the material that they failed to master in the previous year before advancing to more difficult material might improve the performance of low-performing students more dramatically in math than in reading. School systems that implement objective retention policies might consider using math as well as reading tests as determinants of whether a student should be promoted.

One important implication of our results is that students who are given an exemption and promoted despite their failure to demonstrate reading proficiency apparently would have been likely to benefit from another year in the third grade. While students who were subject to the retention program outperformed those who were not regardless of whether they were actually retained, our second analysis shows that low-performing students who were retained significantly outperformed those who were promoted.

This does not mean that it would be wise to eliminate all exemptions to the testing requirement. There are certainly students for whom testing is either inappropriate or whose performance on other academic measures could reasonably indicate that they would be better served by moving on to the next grade. However, our findings do indicate that teachers and school systems should be cautious when granting exemptions and that many of the students who were promoted during the first year of Florida's program would have made greater gains if they had been retained.

This study provides important information about the effects of Florida's retention policy in the year after students are retained. However, the more important question of the policy's effects over time is not answered in our analysis. Further evaluations of the program are needed to follow these same cohorts of students throughout their academic careers to find whether the gains made by students subject to Florida's retention policy grow through accumulation, recede, or remain constant over time. It will also be possible to evaluate whether this retention policy for third-grade students has any substantial effect on other outcomes, including the probability that they will graduate from high school. Further research on similar programs throughout the nation would also prove useful. However, it is reasonable to assume that our results in Florida have strong implications for other school systems, such as Texas, Chicago, and New York, that have implemented similar programs.

The findings of this study demonstrate that after one year, Florida's retention policy has significantly improved the academic proficiency of low-performing third-grade students. Further research on this and other programs will add vital information to the debate over objective retention policies. For now, however, the early results are quite encouraging for the use of retention based on standardized tests to improve academic proficiency.

ENDNOTES

1. See Florida's ranking on the National Assessment of Educational Progress (NAEP), a standardized test administered by the U.S. Department of Education (http://www.nces.ed.gov/nationsreportcard/). Florida also had the lowest graduation rate in the nation for the class of 2001 (see Greene and Forster 2003).

2. See http://info.fldoe.org/dscgi/ds.py/Get/File-434/grade_3_reading_.pdf. In order to receive an exemption under criterion no. 3, the student must score above the fifty-first percentile on the Stanford-9.

3. Authors' calculations using data provided by Florida Department of Education.

4. Authors' calculations using data taken from Table 38 of the *Digest of Education Statistics 2003* and Table 19 of "Characteristics of the 100 Largest Public Elementary and Secondary School Districts in the United States: 2001-02," both published by the National Center for Education Statistics.

5. In order to score above the Level 1 threshold on the FCAT reading test, a student must have a developmental-scale score of at least 1,046. This threshold was the same for groups in both years of our analysis.

6. The norm-referenced version of the FCAT is actually the Stanford-9.

7. Such concerns remain widespread, though previous research indicates that the results of high-stakes tests, particularly the FCAT, are reliable (see Greene, Winters, and Forster 2003) precisely because their results correlate highly with those of low-stakes tests.

8. We were able to obtain information on whether the student was non-Hispanic white, non-Hispanic African-American, Hispanic, Asian, Indian, or multiracial.

9. Eric A. Hanushek, "The Failure of Input-Based Schooling Policies," *Economic Journal*, February 2003.

10. See Jay P. Greene, "The Effect of School Choice: An Evaluation of Charlotte's Children's Scholarship Fund," Manhattan Institute, August 2002. See also William Howell and Paul E. Peterson, "The Education Gap: Vouchers and Urban Schools," Brookings Institution Press, 2002.

REFERENCES

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber, "On the Success of Failure: A Reassessment of the Effects of Retention in the Primary School Grades, Second Edition" Cambridge University Press, 2003.

Greene, Jay P., Marcus A. Winters, and Greg Forster, "Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?" *Teachers College Record*, June 2004, Vol. 106, Num. 6, pp 1124–1145.

Harrison, Steve, "'Unusual' Test Scores Raise Questions: A Herald analysis finds that big gains on the Florida Comprehensive Assessment Test recorded at several schools in Broward and Miami-Dade counties are statistically improbable" *Miami Herald*, February 8, 2004.

Holmes, C. Thomas, "Grade Level Retention Effects: A Meta-Analysis of ResearchStudies" in Eds. Lorrie A. Shepard and Mary Lee Smith "Flunking Grades: Research and Policies on Retention" The Falmer Press, 1989.

Nagoaka, Jenny and Melissa Roderick, "Ending Social Promotion: The Effects of Retention" Consortium on Chicago School Research, March 2004.

Reynolds, Arthur J., "Grade Retention and School Adjustment: An Explanatory Analysis" *Educational Evaluation and Policy Analysis*, Summer 1992, Vol. 14 No. 2, pp 101–121.

Appendix: Tables

Table 1: Gains Made by Students Translated into Standard Deviation Units

	Standard Deviation Unit Gain for Students Subject to Policy Relative to Students Not Subject to Policy	Standard Deviation Unit Gain for Retained Students Relative to Promoted Students
FCAT Reading	0.06	0.13
Stanford-9 Reading	0.06	0.11
FCAT Math	0.15	0.30
Stanford-9 Math	0.14	0.28

Table 2: Gains Made by Students Translated into Percentile Scores

	Percentile Gain for Students Subject to Policy Relative to Students Not Subject to Policy	Percentile Gain for Retained Students Relative to Promoted Students
FCAT Reading	1.85	4.10
Stanford-9 Reading	1.85	3.45
FCAT Math	4.76	9.98
Stanford-9 Math	4.43	9.26

Note: Percentile gain calculated for student with average baseline Stanford-9 reading score among all students in our analysis (23rd percentile)

Table 3: Effect of Being Subject to Retention Policy on FCAT Reading Test

	Effect on FCAT Reading Test	Standard Error	P-Value
Student Is Subject to Policy	16.66	1.92	0.0000
American Indian	1.91	20.09	0.9242
Asian or Pacific Islander	36.56	9.67	0.0002
Black, Not Hispanic	-38.67	2.50	0.0000
Hispanic	-3.48	3.20	0.2768
Multiracial	14.95	7.31	0.0409
Receives Either Free or Reduced-Price Lunch	-54.81	2.42	0.0000
Student Is Limited English Proficient	-2.33	2.87	0.4169
Baseline FCAT Reading Test Score	-0.48	0.00	0.0000
Adjusted R-Squared	0.161		
N	89604		

Table 4: Effect of Being Subject to Retention Policy on Stanford-9 Reading Test

	Effect on Stanford-9 Reading Test	Standard Error	P-Value
Student Is Subject to Policy	1.44	0.17	0.0000
American Indian	-0.90	1.76	0.6103
Asian or Pacific Islander	3.85	0.84	0.0000
Black, Non-Hispanic	-5.90	0.22	0.0000
Hispanic	-1.06	0.28	0.0001
Multiracial	-0.33	0.64	0.6087
Receives Either Free or Reduced-Price Lunch	-4.13	0.21	0.0000
Student Is Limited English Proficient	2.06	0.25	0.0000
Baseline Stanford-9 Reading Test Score	-0.38	0.00	0.0000
Adjusted R-Squared	0.138		
N	873455		

Table 5: Effect of Being Subject to Retention Policy on FCAT Math Test

	Effect on FCAT Math Test	Standard Error	P-Value
Student Is Subject to Policy	41.67	1.45	0.0000
American Indian	-3.73	15.05	0.8044
Asian or Pacific Islander	54.71	7.28	0.0000
Black, Not Hispanic	-29.25	1.90	0.0000
Hispanic	-1.92	2.41	0.4274
Multiracial	1.14	5.51	0.8367
Receives Either Free or Reduced-Price Lunch	-25.77	1.83	0.0000
Student Is Limited English Proficient	7.75	2.16	0.0003
Baseline FCAT Math Test Score	-0.45	0.00	0.0000
Adjusted R-Squared	0.245		
N	89209		

Table 6: Effect of Being Subject to Retention Policy on Stanford-9 Math Test

	Effect on Stanford-9 Math Test	Standard Error	P-Value
Student Is Subject to Policy	4.50	0.16	0.0000
American Indian	0.31	1.71	0.8576
Asian or Pacific Islander	4.59	0.82	0.0000
Black, Not Hispanic	-4.87	0.21	0.0000
Hispanic	-0.71	0.27	0.0088
Multiracial	-0.78	0.63	0.2136
Receives Either Free or Reduced-Price Lunch	-2.44	0.21	0.0000
Student Is Limited English Proficient	0.62	0.24	0.0115
Baseline Stanford-9 Math Test Score	-0.35	0.00	0.0000
Adjusted R-Squared	0.168		
N	873789		

Table 7: Effect of Retention on FCAT Reading Test

	Effect on FCAT Reading Test	Standard Error	P-Value
Student Was Retained	32.48	3.80	0.0000
American Indian	2.44	20.18	0.9040
Asian or Pacific Islander	37.51	9.71	0.0001
Hispanic	-3.66	3.22	0.2555
Multiracial	14.69	7.35	0.0458
Black, Non-Hispanic	-39.65	2.52	0.0000
Receives Either Free or Reduced-Price Lunch	-56.42	2.44	0.0000
Student Is Limited English Proficient	-1.41	2.88	0.6260
Baseline FCAT Reading Test Score	-0.48	0.00	0.0000
Adjusted R-Squared	0.159		
N	89397		

Table 8: Effect of Retention on Stanford-9 Reading Test

	Effect on Stanford-9 Reading Test	Standard Error	P-Value
Student Was Retained	2.80	0.33	0.0000
American Indian	-0.82	1.75	0.6416
Asian or Pacific Islander	3.96	0.84	0.0000
Hispanic	-1.06	0.28	0.0002
Multiracial	-0.45	0.64	0.4849
Black, Non-Hispanic	-6.00	0.22	0.0000
Receives Either Free or Reduced-Price Lunch	-4.27	0.21	0.0000
Student Is Limited English Proficient	2.15	0.25	0.0000
Baseline Stanford-9 Reading Test Score	-0.37	0.00	0.0000
Adjusted R-Squared	0.138		
N	87163		

Table 9: Effect of Retention on FCAT Math Test

	Effect on FCAT Math Test	Standard Error	P-Value
Student Was Retained	82.54	2.87	0.0000
American Indian	-2.59	15.05	0.8635
Asian or Pacific Islander	56.56	7.28	0.0000
Hispanic	-2.14	2.42	0.3768
Multiracial	1.16	5.51	0.8339
Black, Non-Hispanic	-31.22	1.90	0.0000
Receives Either Free or Reduced-Price Lunch	-29.11	1.83	0.0000
Student Is Limited English Proficient	9.73	2.17	0.0000
Baseline FCAT Math Test Score	-0.44	0.00	0.0000
Adjusted R-Squared	0.244		
N	89007		

Table 10: Effect of Retention on Stanford-9 Math Test

	Effect on Stanford-9 Math Test	Standard Error	P-Value
Student Was Retained	8.77	0.32	0.0000
American Indian	0.40	1.69	0.8136
Asian or Pacific Islander	4.77	0.81	0.0000
Hispanic	-0.77	0.27	0.0044
Multiracial	-0.82	0.62	0.1845
Black, Non-Hispanic	-5.08	0.21	0.0000
Receives Either Free or Reduced-Price Lunch	-2.78	0.20	0.0000
Student Is Limited English Proficient	0.85	0.24	0.0004
Baseline Stanford-9 Math Test Score	-0.33	0.00	0.0000
Adjusted R-Squared	0.171		
N	87188		

The Center for Civic Innovation's (CCI) purpose is to improve the quality of life in cities by shaping public policy and enriching public discourse on urban issues.

CCI sponsors the publication of books like The Entrepreneurial City: A How-To Handbook for Urban Innovators, which contains brief essays from America's leading mayors explaining how they improved their cities' quality of life; Stephen Goldsmith's The Twenty-First Century City, which provides a blueprint for getting America's cities back in shape; and George Kelling's and Catherine Coles' Fixing Broken Windows, which explores the theory widely credited with reducing the rate of crime in New York and other cities. CCI also hosts conferences, publishes studies, and holds luncheon forums where prominent local and national leaders are given opportunities to present their views on critical urban issues. *Cities on a Hill*, CCI's newsletter, highlights the ongoing work of innovative mayors across the country.

The Manhattan Institute is a 501(C)(3) nonprofit organization. Contributions are tax-deductible to the fullest extent of the law. EIN #13-2912529

MANHATTAN INSTITUTE FOR POLICY RESEARCH

52 Vanderbilt Avenue • New York, NY 10017
www.manhattan-institute.org