Getting Ahead by Staying Behind

An Evaluation of Florida's Program to End Social Promotion

Of the many entrenched school customs that have been reconsidered and reformed over the past decade, social promotion has been among the most resistant to change. Holding children back in the same grade has long been frowned upon, and a large body of research seems to support that point of view: retained students tend to have lower test scores and are allegedly more likely to drop out than students who initially performed at an equally low level but were nevertheless promoted.

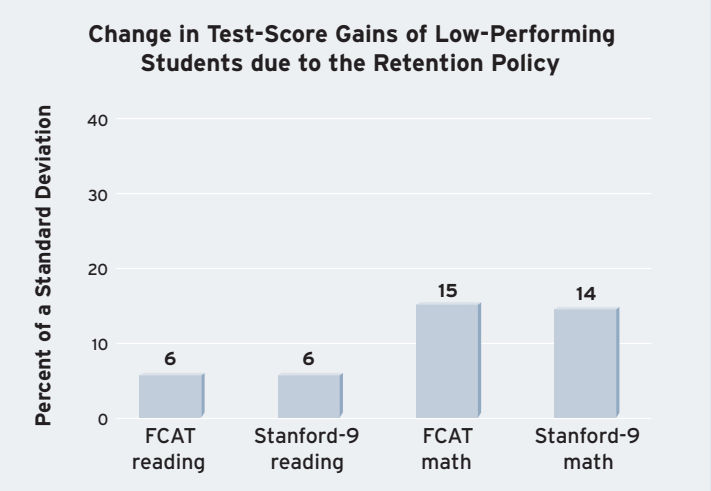
Despite the old habits and the old research, however, school districts across the nation have been slowly but steadily bucking convention. Several large systems, including Chicago (beginning in 1996), New York (2004), and Philadelphia (2005), now require students in particular grades to demonstrate a benchmark

PHOTOGRAPH / GETTY IMAGES

BY JAY P. GREENE AND MARCUS A. WINTERS

A Productive Policy (Figure 1)

Low-performing 3rd graders subjected to Florida's new retention policy in 2003 made larger test-score gains the following year than did comparable students entering 3rd grade in 2002.



Note: All effects are statistically significant at the 0.001 level and control for differences in race, free or reduced-price lunch status, Limited English Proficiency status, and prior test scores.

SOURCE: Authors' calculations from Florida Department of Education data

level of mastery in basic skills on a standardized test before they can be promoted. Florida (2002) and Texas (2002) have taken the lead among states in forbidding social promotions. In 2000, the most recent year for which national enrollment data are available, these five school systems alone enrolled nearly 20 percent of the nation's 3rd-grade students. (For more on Chicago's policy, see Alexander Russo, "Retaining Retention," *features*, Winter 2005; and Robin Tepper Jacob and Susan Stone, "Teachers and Students Speak," *features*, Winter 2005.)

But is this new approach to grade promotion effective? And what about those studies that say retention doesn't work? Proponents of the new programs believe that schools do students no favor by promoting them if they don't have the skills to succeed at a higher level. But because these arguments, however plausible, have little research to support them, we set out to determine if they have scientific merit. Our findings from Florida suggest that the use of standardized testing policies to end social promotion can help low-performing students make modest improvements in reading and substantial improvements in math.

Florida's Program to End Social Promotion

Over the past several years Florida has attempted substantial reforms of its struggling public school system, the fourth-largest in the country and one that consistently ranks close to the bottom on academic indicators, including high-school graduation rates and scores on the National Assessment of Educational Progress (NAEP). The Sunshine State had instituted school voucher programs, increased the number of charter schools, and devised a sophisticated accountability system that evaluates schools on the basis of their progress as measured by the Florida Comprehensive Assessment Test (FCAT). But in May 2002, the state legislature made one of its boldest moves, revising the School Code, the state's education law, to require 3rd-grade students to score at the Level-2 benchmark or above on the reading portion of the FCAT in order to be promoted to 4th grade.

The hurdle created for students was not terribly high. The state's department of education describes a student who scores at Level 2 (of five levels) as having "limited success" against the state standards; only students who score at Level 3 or above are considered to be proficient for the purposes of evaluating schools under No Child Left Behind. Even so, roughly 24 percent of 3rd graders tested in Florida in 2001–02, the year before the retention policy was introduced, performed at Level 2 or below. This number fell slightly, to 22 percent, in the 2002–03 academic year.

Not all these students were retained, however, even after the policy change. The law allowed for exceptions to the retention policy if a student had limited English proficiency or a severe disability, scored above the 51st percentile on the Stanford-9 standardized test, had demonstrated proficiency through a performance portfolio, or had already been held back for two years. Altogether, roughly 40 percent of the 3rd-grade students who scored below the Level-2 threshold in 2002–03 were promoted.

The Problem with Earlier Studies

Traditionally, the retention of a student, uncommon as it was, resulted from an individual teacher's assessment of the student's ability to succeed at the next level. But such teacher discretion, while arguably desirable as a matter of policy, is the primary reason earlier studies of social promotion are flawed. We must assume from studying those retention programs, which are still the predominant practice in schools throughout the United States, that students who were held back were fundamentally different from students who were promoted. Because teachers were considering intangible factors, even when race, gender, family income, and academic achievement are the same, there was no way to isolate the effect of being held back, much less to make reasonable conclusions about the effects of retention on a student's academic achievement or the probability of his dropping out

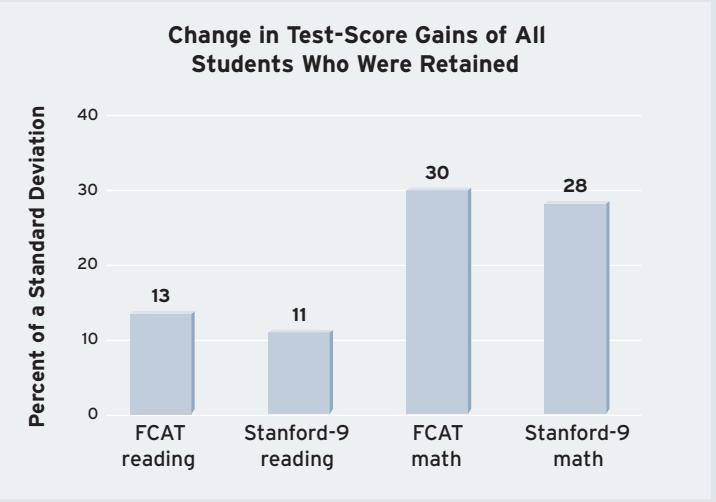
of high school. Are students who were retained less likely to graduate because they were retained? Or were they retained because of characteristics that also predisposed them to drop out? Because the retention policies were subjective, we will simply never know.

There are also reasons to believe that subjective retention policies affect students differently than policies that use promotion criteria like performance on standardized tests. If promotion depends on an individual teacher's assessment of a child, then that child is not likely to know what he or she must do to avoid being held back. Also, if few students were being held back, then those students might perform worse because they felt excluded and inferior. A policy that holds back thousands of students might dilute this sense of being singled out. Finally, 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 inhibit.

Implementing objective standards, even if they were accompanied by subjective exemptions, might significantly change the effects of retention in ways that previous research could not anticipate or measure. For research purposes, objective retention policies also create a useful comparison group of students not subject to retention. In the case of Florida's program to end social promotion, for example, we can compare students who were subject to the threat of retention with students who would have been had they been born a year later.

Retention Works (Figure 2)

Students retained in 2003 as a result of the new policy made substantially more progress in reading and, especially, in math than comparable students who were promoted.



Note: All effects are statistically significant at the 0.001 level and are adjusted for differences in race, free or reduced-price lunch status, Limited English Proficiency status, and prior test scores.

SOURCE: Authors' calculations from Florida Department of Education data

What a Difference a Year Makes

To determine the impact of ending social promotion for 3rd graders in Florida, we compared low-scoring 3rd graders in 2002, the first students to be subject to the program, with low-scoring 3rd graders from the previous year. Of the 43,996 3rd graders in 2002 for whom we have valid test scores on both FCAT math and reading assessments, 60 percent were actually retained. By contrast, of the 45,401 3rd graders in 2001 for whom we have valid test scores, only 9 percent were retained. Our analysis assumes that the students from the two school years should be similar in all respects except for the year in which they happened to have been born. We analyzed the test-score improvements made between each student's first 3rd-grade year and the following year on both the state's own accountability exam and the Stanford-9, a nationally normed exam administered at the same time as the FCAT but not used for accountability purposes.

We measure FCAT performance using developmental-scale scores, which allow us to compare the test-score gains of all the students in our study, even though they took 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 3rd grader with a developmental-scale score of 1,000 and a 4th 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 and then gets the same score of 1,000 in 2002, this indicates that the student has not made any academic progress in the intervening year. The developmental-scale scores required to reach Level 2 on the FCAT reading test were consistent for each year's cohort.

We began by measuring the effect on all low-scoring 3rd graders of simply having been subject to the new policy. That is, we did not distinguish in our initial analysis between students who were actually retained and those who received an exemption and were promoted to the next grade. This analysis provides an estimate of the average impact of the policy change on all students in the state performing below the Level-2 benchmark. It also allows for the possibility that exempted students enjoyed spillover benefits from the retention policy, since they were now being instructed in a system in which fewer students in 4th grade were unprepared to do grade-level work.

To identify the policy's average impact, we compared the gains in developmental-scale scores made by students who first entered 3rd grade in 2002 and scored below the FCAT

benchmark with gains made by students who first entered 3rd grade in 2001 and scored below the FCAT benchmark. In making this comparison, we took into account other factors that could affect achievement gains, such as the student's race, whether the student received a free or reduced-price school lunch, whether the student was deemed Limited English Proficient, and the student's precise test score during his first 3rd-grade year. With these differences accounted for, the only distinction between the two groups of students was assumed to be that the former group entered the school system a year later and was therefore subject to the new policy in 3rd grade.

As discussed above, however, many low-scoring 3rd graders were granted exemptions and promoted to the 4th grade even under the new policy. We therefore also evaluated the effect of actually being retained, again controlling for race, eligibility for free or reduced-price lunch, English proficiency, and baseline test scores. In conducting this analysis, we also needed to account for the fact that the students who were held back were a select group of students who could differ in important ways from the promoted students. Presumably, teachers and other decisionmakers expected these students, unlike promoted students, to benefit from an additional year as 3rd graders. Fortunately, the fact that simply having entered school a year later increased the probability of retention for all low-scoring students again provides a way around this obvious selection problem. In essence, the statistical method we use compares those retained students that our data suggest would not have been retained the previous year with a comparable group of students who were not retained. Our results therefore indicate the effect of retention on those students who were held back as a result of the new policy.

During this time, Florida was engaged in other education reforms as well: instituting several school-voucher programs, increasing the number of charter schools in the state, and improving the system used to assign grades to schools based on the FCAT. However, it is reasonable to assume that whatever effect these other policies have on our analyses is minor. In order for the existence of another policy to affect our results significantly, we would have to believe that the program substantially improved the education of the 3rd graders in 2002–03 without having a similar effect on the previous year's cohort. Moreover, while a sudden policy change could conceivably explain the overall improvements between the two cohorts, it is difficult to see how such a change could cause substantially larger gains among those students actually retained.

Retention Works

Our fundamental findings from an analysis of the 3rd- and 4th-grade data for these two years indicate that the performance of students identified for retention, regardless of whether they were retained or exempted and promoted, exceeded the performance of low-performing students from the previous year who were not subject to the retention policy; and students who were actually retained made the larger relative gains.

Students identified for retention by the Florida policy gained 0.06 of a standard deviation in reading on both the FCAT and Stanford-9 over equally low-performing 3rd graders from the previous school year (see Figure 1). In math, students identified for retention surpassed low performers who were not subject to the policy by 0.15 standard deviations (4.8 percentiles) on the FCAT and 0.14 standard deviations (4.4 percentiles) on the Stanford-9.

Students who were actually retained experienced even larger relative improvements (see Figure 2). Retained students performed better than low-scoring students who were promoted by 0.13 standard deviations (4.10 percentiles) on the FCAT and 0.11 standard deviations (3.45 percentiles) on the Stanford-9 in reading. In math retained students improved 0.30 standard deviations (10.0 percentiles) on the FCAT and 0.28 standard deviations (9.3 percentiles) on the Stanford-9 over promoted students.

Some critics of the new retention policies argued that teachers and schools would respond to them by manipulating test scores, either directly by cheating or indirectly by teaching students skills that would help them to improve their test scores but would not provide real academic proficiency. This argument would have merit only if we found strong gains on the high-stakes FCAT and no similar gains on the low-stakes Stanford-9, for which there is no incentive to manipulate scores. But our results are consistent between the FCAT and the Stanford-9, indicating that there have been no serious manipulations of the high-stakes testing system. 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 would indicate that teachers have made changes resulting in real increases in students' proficiency.

An unexpected benefit of the retention policy is the improvement in math scores. This might seem 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. Of course, the math gains could simply reflect the fact that math skills are learned primarily in schools, while reading is practiced both in and outside of school. For this reason, evaluations of school reforms frequently find stronger effects in math than in reading. Alternatively, it may be that students who were retained specifically because of their poor reading skills are particularly poor in that subject and that this limits their room for improvement.

We also explored the possibility that the objective retention program could have different effects on students of different races. Our results show gains of similar sizes by the three racial groups for which we have an adequate sample size to have reasonable confidence in our findings: white, black, and Hispanic. The exception is for whites' performance on the FCAT reading test. It is difficult for us to interpret why white students would fail to benefit from the retention policy as measured by the FCAT reading test but would be shown to benefit as measured by the Stanford-9 reading test.

Our results show gains of similar sizes by the three racial groups for which we have an adequate sample size: white, black, and Hispanic.

Our results also suggest that low-scoring Florida 3rd graders who were given an exemption and promoted might have benefited from another year in the 3rd grade. 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.

What It Means

At first glance our findings seem inconsistent with evaluations of Chicago's program ending social promotion, to our knowledge the only similarly designed retention policy to be evaluated using comparable methods. In Chicago, students in the 3rd, 6th, and 8th grades must exceed benchmarks on the Iowa Test of Basic Skills (ITBS), a respected standardized test, in order to be promoted to the next grade. In a study conducted in 2004 by scholars at the Consortium on Chicago School Research, the performance of 3rd- and 6th-grade students who scored just below the benchmark on the ITBS, most of whom were retained because of the mandate, was compared with the performance of students who scored just above the benchmark, most of whom were promoted. The Chicago researchers were able to measure test-score performance for two years after implementation of the program. They found benefits from the program after one year, similar to what we found in Florida, but discovered that those benefits went away after the second year. Third-grade students were not affected, and 6th-grade students were negatively affected by the policy in their performance on the ITBS reading test. The findings on the Chicago retention program emphasize the importance of following the progress of retained students in Florida over time.

Still, the Chicago policy differs from Florida's in some respects. In 1999 the Chicago policy stopped allowing students to be retained twice, which Florida's policy does allow. This difference might reduce teachers' motivation to work with already retained students, whom they now can expect to be promoted the next year regardless of their performance. Other programs with different and more stable retention policies might show different results.

Finally, while our study provides valuable information about the effectiveness of Florida's policy to end social promotion, it does not offer a full catalog of the policy's benefits or of its potential costs. It will be some time before we can examine whether retention increased or reduced the probability of dropping out of school later on. Most important, it does not provide any information about the program's effects on students' academic progress the first time they were in 3rd grade. The policy's greatest benefits could result not from retention itself, but rather from increased efforts on the part of teachers and even students to avoid being retained in the first place.

Jay P. Greene is professor and head of the Department of Education Reform at the University of Arkansas; he is also a senior fellow at the Manhattan Institute. Marcus A. Winters is a doctoral fellow at the University of Arkansas and a senior research associate at the Manhattan Institute.