Elementary- and secondary-school teachers in the United States traditionally have been compensated according to salary schedules based solely on experience and education. Concerned that this system makes it difficult to retain talented teachers and provides few incentives for them to work to raise student achievement while in the classroom, many policymakers have proposed merit-pay programs that link teachers’ salaries directly to their apparent impact on student achievement.

Until recently, only a handful of isolated districts had attempted such programs. Now entire state systems are moving toward merit pay, with new policies established recently in Florida and Texas requiring districts to set teachers’ salaries based in part on the gains their students are making on the state’s accountability exam.

Implementing a merit-pay system, however, comes with challenges. Students often have more than one teacher but take only one high-stakes test. How do we know which teacher to reward? If students...
are not tested annually in each subject, how do we determine the merit of a teacher in a year without testing? How do we fairly assess the impact of a teacher during a testing year if we do not know how students performed during the previous school year? Can a merit-pay system overcome these obstacles?

One option is to turn to principals and ask them to help determine the size of pay raises. Such subjective performance assessments are already used to evaluate untenured teachers, and they play a large role in promotion and compensation decisions in other occupations. While principals can and do judge teachers’ performance, however, there is little good evidence on the accuracy of their judgments.

The research reported in this paper fills this gap. We found that principals in a western school district did a good job of assessing teachers’ effectiveness. In fact, principals are quite good at identifying those teachers who produce the largest and smallest standardized achievement gains in their schools (the top and bottom 10–20 percent). They are less able to distinguish among teachers in the middle of this distribution (the middle 60–80 percent), suggesting that merit-pay programs that reward or sanction teachers should be based on evaluations by principals and should be focused on the highest- and lowest-performing teachers.

**A Representative Sample**

We surveyed all 13 elementary-school principals in a mid-sized school district, that asked to remain anonymous, in the western United States. We asked them to rate the teachers in their schools on a variety of performance dimensions. The survey, conducted in February 2003, provides evaluations by their principals of 202 elementary-school teachers in grades 2 through 6.

The teachers included in the study are fairly representative of elementary-school teachers nationwide. Sixteen percent of them are men, the average age is 42, and average teaching experience is 12 years. Most of these teachers attended a local university; 10 percent attended another in-state college; and 6 percent attended a school out of state. Seventeen percent of them have a master’s degree or higher, and most are licensed in either early childhood education or elementary education. Finally, 8 percent of the teachers in our sample taught in a mixed-grade classroom in 2002–03, and 5 percent were in a “split” classroom, sharing a single contract and dividing the school day with another teacher. The students in grades 2 through 6 in the district are predominantly white (73 percent), with a sizable ethnic minority (Latino students compose 21 percent of the elementary population); 48 percent of them receive a free or reduced-price lunch. Achievement levels in the district are almost exactly at the average of the nation (49th percentile on the Stanford Achievement Test).

All elementary-school students in the district take a set of exams each year, in reading and math. These multiple-choice, criterion-referenced tests cover topics that are closely linked to the district’s learning objectives. While student achievement results have not been linked to rewards or sanctions for schools until recently, the results of the exams have been distributed to parents annually for at least the past decade, years before implementation of the No Child Left Behind law. This latter fact is important because our study relies on a consistent data set covering the years 1998 through 2003. The district has not had a merit-pay program for teachers at any time during this period.

To ensure that we could link student achievement data to the appropriate teacher, we limited our sample to classroom teachers, omitting music and gym teachers as well as librarians. We excluded kindergarten and first-grade teachers because earlier achievement exams were not available for their students; this prevented us from developing a “value-added” measure of student learning. We retain in our analysis the small number of teachers who share a contract, each teaching only half of the school day. For our analysis, the gains made by students in these classes count toward the estimated value added of each of the two teachers.

**Can Principals Identify Effective Teachers?**

Principals were asked not only to provide a rating of overall teacher effectiveness, but also to assess, on a scale from one (inadequate) to ten (exceptional), specific teacher characteristics (ten altogether), including dedication and work ethic, classroom management, parent satisfaction, positive relationship with administrators, and ability to improve math and reading achievement. Principals were assured that their responses would be completely confidential and would not be revealed to the teachers or to any other employee of the school district.

While there was some variation among principals, the overall assessments they gave teachers were generally quite high, with an average of 8.1. Only 10 percent of the assessments fell below a 6, and the average rating for the least-generous principal was still a 6.7. At the same time, principals did not simply assign similar scores to each of their teachers. In fact, the principals generally used 5 to 6 different ratings for the teachers in their school.

Because principals differ in the generosity and degree of variation in the ratings they give, we placed all the ratings on the same scale by subtracting from each teacher’s rating the average rating given by that teacher’s principal and then dividing by the principal’s standard deviation. We did this separately for each specific aspect of teacher performance about which principals were asked.

We compared a principal’s assessment of how effective a teacher is at raising student reading or math achievement,
one of the specific items principals were asked about, with that teacher’s actual ability to do so as measured by their value added, the difference in student achievement that we can attribute to the teacher. To estimate the value added by a teacher, we examine the performance of her students after accounting for a wide variety of student and classroom characteristics that could affect achievement independent of the teacher’s ability. These characteristics include race, gender, eligibility for the federal lunch program, limited English proficiency, and, most important, previous student achievement. We also take advantage of the availability of data on the same teachers from as far back as the 1996–97 school year; this enables us to distinguish long-term teacher quality from the possibly idiosyncratic performance of a class in any one year.

We find a positive correlation between a principal’s assessment of how effective a teacher is at raising student achievement and that teacher’s success in doing so as measured by the value-added approach: 0.32 for reading and 0.36 for math. These correlations are based not on a principal’s overall rating of the teacher, but rather on the principal’s personal assessment of how effective the teacher is at “raising student math (or reading) achievement.” Previous studies of evaluations by principals have used only the overall rating of the teacher, a less direct assessment of a teacher’s ability to raise student performance. Using the overall rating in that way could compromise the accuracy of subjective performance evaluations, especially if principals value characteristics of teachers that are unrelated to their effect on student performance. Our findings lead us to conclude that principals are able to identify accurately this dimension of teacher effectiveness.

Why aren’t these correlations even higher? One possible explanation is that principals focus on the average test scores in a teacher’s classroom rather than on student improvement. There is some evidence for this conjecture. The correlation between ratings by principals and the average test scores of a teacher’s students is significantly higher than the correlation between ratings by principals and the teacher’s value-added rating in reading (0.56 versus 0.32), though not in math.

Another reason could be that principals focus on their most recent observations of teachers. We do find, for example, that the average achievement gains in a teacher’s classroom in 2002–03 is a modestly stronger predictor of the principal’s rating than the gains in any previous year. In theory, it is possible that principals are correct in assuming that a teacher’s effectiveness changes over time so that teachers’ most recent experience is the best indicator of their actual effectiveness. If that were the case, however, we would expect to find that principals’ ratings are more highly correlated with value-added measures that have been adjusted to account for the fact that teachers tend to be less effective in their first one or two years in the classroom. In fact, the correlation between principals’ ratings and experience-adjusted value-added measures is no higher than the correlation with our baseline value-added measures. The bigger mistake principals make, it seems, is not adequately accounting for students’ incoming ability.

While informative about principals’ overall abilities, a simple correlation does not tell us whether principals are more or less effective at identifying teachers at certain points on the ability distribution. We therefore estimated the percentage of teachers that a principal can correctly identify in the top group within his or her school. We found that the teachers identified by principals as being in the top category were, in fact, in the top category according to the value-added
measures about 52 percent of the time in reading and 69 percent of the time in mathematics. If principals randomly assigned ratings to teachers, we would expect the corresponding probabilities to be 14 and 26 percent, respectively. This suggests that principals have considerable ability to identify teachers in the top of the distribution. The results are similar if one examines principals' ability to identify teachers in the bottom of the ability distribution.

Principals do a reasonably good job of identifying those teachers who are better (and worse) at raising student test scores. Not surprisingly, the best way to predict how effective a teacher will be is to find out how effective the teacher has been in the past. Differences in teachers' salaries within a school system are entirely unrelated to teachers' effectiveness.

Despite their success with the top and bottom of the distribution, principals are significantly less successful at distinguishing among teachers in the middle of the ability distribution. Principals correctly identify only 49 percent of teachers as being better than the median teacher in their school in boosting students' reading scores, relative to the 33 percent that one would expect if ratings were randomly assigned. Principals appear somewhat better at distinguishing between teachers in the middle of the distribution in math (they correctly placed 54 percent of teachers above the median, compared with the 26 percent expected if ratings were random), but they again appear to be better at identifying the best and worst teachers.

One reason that principals might have difficulty distinguishing between teachers in the middle is that the distribution of teachers’ value-added ratings is highly compressed. However, our analysis of the data suggests that this is not the case. Teachers who receive ratings at or close to the median in the school have estimated value-added measures that are quite widely dispersed.

**What Characteristics of Teachers Do Principals Value?**

Of course, the effects of moving to a system of compensation based on assessment by principals depend on the relative importance they place on a teacher’s ability to raise standardized test scores when making overall assessments of teachers' effectiveness. While such preferences could theoretically be set by district administrators or other policymakers, it is likely that principals would retain some autonomy over personnel decisions, so their preferences are important to investigate. We therefore compared principals’ overall rating of each teacher with their assessment of various teacher attributes to examine how principals value different dimensions of quality in teachers.

Perhaps not surprisingly, teachers’ ratings on many (though not all) of the individual survey items are highly correlated. Based on the relationships between the questions, we created three groups of teachers’ quality characteristics and reanalyzed the results. The first group captures what might be described as traditional teaching ability and includes the ratings of classroom management, organization, and ability to improve students’ test scores. The second, including the principal’s assessments of a teacher’s relationship with colleagues and administrators, measures a teacher’s collegiality. The third measures student satisfaction and includes the principal’s ratings of student satisfaction and the teacher as a role model.

Ability, collegiality, and student satisfaction all contribute independently to a principal’s overall evaluation of a teacher,
but principals weigh the set of questions measuring teachers’ ability to improve student achievement and to manage a classroom most heavily. An increase of one standard deviation in a principal’s evaluation of a teacher’s management and teaching ability, for example, is associated with an increase of 0.56 standard deviations in the principal’s overall rating. In comparison, an increase of one standard deviation in teacher collegiality is associated with an increase in overall ratings of roughly one-third of a standard deviation in overall rating. Meanwhile, teachers scoring one standard deviation higher in student satisfaction score just 0.15 standard deviations in their overall rating, all else being equal.

Predicting Performance
We should care about the quality of principals’ assessments of teacher quality not just for their reliability in a merit-pay system, but also for their ability to identify teachers who will continue to improve student achievement. In order to get a sense of how well principals’ assessments forecast teachers’ performance, we examined how well these assessments predict future student achievement gains. For our February 2003 survey of principals, that meant evaluating scores on the spring 2003 tests. We compared the predictive accuracy of a principal’s assessment of teacher effectiveness with the predictive accuracy of a teacher’s value-added rating. We also measured the accuracy of the traditional determinants of teachers’ salaries, experience and education, in predicting those scores. Throughout, we accounted for differences in previous student achievement, student demographics, and classroom characteristics.

Our findings suggest that ratings by principals, both overall ratings and ratings of a teacher’s ability to improve achievement, effectively predict a student’s future achievement gains (see Figure 1). Students whose teachers receive an overall rating one standard deviation above the mean are predicted to score roughly 0.06 standard deviations higher in reading than students whose teacher received an average rating. By way of comparison, students receiving free or reduced-price lunch in the same district experience achievement gains approximately 0.16 standard deviations lower than similar students who are not eligible for such programs. Assignment to a teacher with a favorable evaluation by her principal appears to be more important for math performance. An increase of one standard deviation in the principal’s evaluation predicts an increase of 0.14 standard deviations in math performance, roughly on par with the disadvantage associated with coming from a low-income family.

Measures of teachers’ value added in previous years are an even better predictor of future gains in students’ achievement than are principal ratings. These results, which are similar for math and reading, suggest that teachers’ impact on student achievement, as measured by simple value-added measures of teacher effectiveness, remain fairly stable over time and that principals’ ratings effectively capture a substantial fraction of these stable differences in teachers’ effectiveness.

We do not find any statistically significant relationship between the number of years a teacher has taught and students’ achievement, though this is probably due to the necessary omission of first-year teachers (because we cannot measure their value added for a previous school year). Other studies have found that first-year teachers tend to perform worse on average than experienced teachers. Education does have some predictive power. Teachers with advanced degrees...
have students who score roughly 0.10 standard deviations higher. We hesitate to say that education itself is producing these gains, because a teacher’s level of education is likely to be associated with personal characteristics not accounted for in our analysis, and these may be the very factors responsible for the improvements in student achievement.

Perhaps our most interesting finding is that the salaries teachers in this district received in 2002–03 bore no relation at all to their impact on student achievement. Students with highly paid teachers made no more progress than those with teachers who had low salaries.

Conclusions
In sum, our results suggest that student achievement (as measured by standardized test scores) would probably improve more under a system based on principals’ assessments than in systems where compensation is based solely on education and experience. This is because principals would be able to identify and reward the very best teachers while, at the same time, identifying the least competent teachers for remediation or dismissal.

To the extent that the most important staffing decisions involve sanctioning incompetent teachers and rewarding the very best teachers, a principal-based assessment system may affect achievement as positively as a merit-pay system based solely on student test results. Moreover, evaluation by the principal has the potential to offset some of the potential negative consequences of test-based accountability systems. If principals can observe inputs as well as outputs, they may be able to ensure that teachers increase student achievement through improvements in pedagogy, classroom management, or curriculum rather than teaching to the test. Principals can also evaluate teachers on the basis of a broader spectrum of educational outputs in addition to test scores that parents may value. At the same time, the inability of principals to distinguish between a broad middle range of teacher quality suggests caution in relying on principals for fine-grained performance determinations, as might be required under certain merit-pay policies.

Two important caveats to consider when interpreting our results. First, we conducted our analysis in a context where principals were not being evaluated on the basis of their ability to identify effective teachers. It is possible that principals’ ability to identify the best-performing teachers would be enhanced by a school system where the principals had more responsibility for monitoring teachers’ effectiveness. At the same time, social or political pressures might make principals less willing to assess teachers honestly if their judgments directly influenced teachers’ compensation. Second, our analysis focuses on the source of the teacher assessment; we do not address the type of rewards or sanctions associated with teacher performance. This is clearly an important dimension of any performance management system, and one would not expect either a principal-based or a test-based assessment system to have a substantial impact on student outcomes unless it were accompanied by meaningful consequences.

Brian Jacob is assistant professor of public policy at the John F. Kennedy School of Government, Harvard University and a faculty research fellow with the National Bureau of Economic Research. Lars Lefgren is assistant professor of economics, Brigham Young University.
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.

**Getting Ahead by Staying Behind**

An Evaluation of Florida’s Program to End Social Promotion

BY JAY P. GREENE AND MARCUS A. WINTERS
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 lower-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 below Level 2. 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.

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

---

**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 not retained.
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 otherwise have been promoted students.

Students identified for retention by the Florida policy 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.

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.
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 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, the University of Arkansas; he is also a senior fellow at the Manhattan Institute. Marcus A. Winters is a doctoral fellow, the University of Arkansas and a senior research associate at the Manhattan Institute.
Jump-starting the Charter School Movement:
A Guide for Donors
The Roundtable commissioned this guide in order to provide donors of all shapes and sizes with strategic ways to support a strong charter school movement in their communities and across the nation. This book draws on the experience of the most active and effective charter funders.

A Donor’s Guide to School Choice
By Brian C. Anderson
Several powerful tools for change fall under the term “school choice:” private scholarships, charter schools, tax credits, and vouchers. This guide provides a road map to all these options, with advice from veteran donors on the pros and cons of each. Invaluable for any donor seeking to improve all schools and to help low-income children attend high-quality schools.

These definitive guides are available free to qualified donors.