RESEARCH BRIEF
Research Services

Vol. 0706
March 2008

Dr. Terry Froman, Research Services
Shelly Brown, Research Services
Angela Luzon-Canasi, Assessment & Data Analysis

Third-Grade Retention: A Four Year Follow-Up

Abstract:
This study duplicated the procedures used by Greene and Winters (2006) on data from the Miami-Dade school system with the advantage of an additional two year's worth of information. The results indicated that the effects of the retention policy are far from clear and arguably negative. There is considerable evidence to suggest that the apparent gains of the retained students may have been short-lived if not completely illusory. The lack of precise measurement and a precisely appropriate comparison group prevent an indisputable interpretation. The superficially obvious benefit of retention to some students and the equally obvious detriment of retention to others will likely keep large-scale test-based promotion policies a matter of heated debate subject to political fashion for the foreseeable future.

Beginning in the 2002-03 school year, the revised Florida School Code required third-grade students to demonstrate reading proficiency by scoring at level 2 or higher on the Reading portion of the Florida Comprehensive Assessment Test (FCAT). Students scoring at Level 1 were retained in third grade for another year, unless exempted from mandatory retention for special circumstances. As this rule has affected and continues to affect thousands of third-graders and their families in Miami-Dade County Public Schools (M-DCPS) over the years, it is important to get as clear a picture as possible of the academic effects of this policy.

Greene and Winters (2006) published a second-year follow-up evaluation of the effects of Florida's statewide test-based third-grade retention policy on student achievement. They concluded that

> "after two years of the policy, retained Florida students made significant reading gains relative to the control group of socially promoted students. These academic benefits grew substantially from the first to the second year after retention. That is, students lacking in basic skills who are socially promoted appear to fall farther behind over time, whereas retained students appear to be able to catch up on the skills they are lacking."

Discretionary vs. Test-Based Retention

Earlier research on the academic impact of retention, back when retention was uncommon and based primarily on the good judgment of educators, routinely concluded that retaining students was not beneficial and could even lead to considerable academic distress (Holmes and Mathews 1984, Holmes 1989, Jimerson 2001). Greene and Winters (2006) correctly point out that the conclusions from studies of this kind of "discretionary retention" do not easily generalize to large-scale test-based retention practices. The administrative, social, and academic circumstances of old-style discretionary retention were much different from this new world of test-based retention. Indeed, in Miami-Dade County alone we went from retaining a few hundred third-graders per year before requiring passing scores for promotion to retaining a few thousand third-graders per year after mandatory retention was invoked. Since the basis for retention has changed, the effects on the type of students retained are unlikely to be comparable to interpretations from earlier studies.

Greene and Winters (2006) go on to say that these studies were plagued by the difficulty to find an appropriate control group against which retained students could be compared. Despite a respected procedure attempting to deal with this complaint, the problem of finding an appropriate control group was not completely overcome by the Greene and Winters methodology and it continues to be a dilemma for current studies of test-based retention.

The Analysis Groups

The first cohort group of interest in this study is the group of third-grade students who were retained in the first year of the test-based retention policy. This group is operationally defined as students at M-DCPS in third grade in 2002-03 who were also in third grade in 2003-04. This group, referred to as experimental group 1 (EXP1), consisted of 5,659 students representing 22.5% of the entire 2002-03 third grade cohort.

If the intention is to study the effects of retention, the ideal comparison for this experimental group would be a random subset of students in that same year third grade class who were identified for retention on the basis of their FCAT scores but promoted specifically to provide an appropriate control group for the research design. Unfortunately for researchers, no such group exists. Strictly speaking, anything short of this is a compromise in research design principles that opens the door to alternate explanations of results. Fortunately, there are a few such compromise control groups that are similar enough in critical dimensions that the research community can feel relatively assured in their interpretations.

Following the lead of Greene and Winters (2006) in what they refer to as an "across-year comparison," a control group of similarly low-scoring students who were not subject to the retention policy because they entered into the system a year before its invocation can be identified. This group is here operationally defined as students at M-DCPS in third grade in 2001-02 (the previous year of the retention policy) who also scored at Level 1 on the FCAT Reading test. This control group 1 (CON1) consists of 9,834 students representing 35.1% of the entire 2001-02 third grade cohort.

The Achievement Outcome Scores

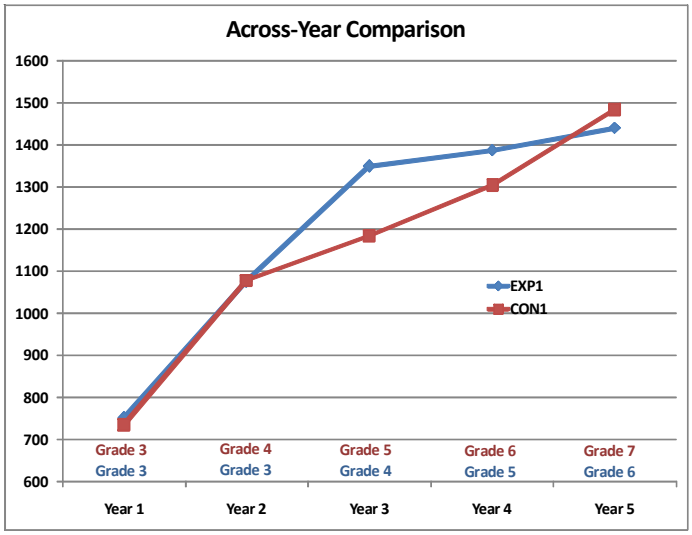
In addition to the FCAT scores reported in a separate scale for each grade level, the Florida Department of Education produces Developmental Scale Scores that are intended to provide for a uniform measure of proficiency across grade levels and years. Using the Developmental Scale, for example, a student earning a score of 1000 on the third-grade FCAT Reading test in 2002-03 may be interpreted as having the same proficiency as a fourth-grade student earning a score of 1000 on the Reading test in 2003-04.

To provide a single scale that yields equivalent measures of academic proficiency across different tests, grade levels, calendar years, and student groups is an ambitious goal. Despite laudable efforts to rigorously develop such a scale by the State, anomalies in the grade level slopes and adaptations to the test over time raise questions as to the Developmental Scale's dependability. In its defense, research exists that supports the FCAT Developmental Scale's validity by producing similar results to those of other standardized tests (Greene, Winters, and Forster 2004; West and Peterson 2005). However, the inevitably inexact nature of

academic achievement measurement suggests caution in interpreting results, particularly in this kind of across-year comparison.

Across-Year Comparison Results

The graph below depicts the average Developmental Scale scores for the two student groups across the five years in this study. Again, the experimental group EXP1 is made up of students retained in third grade and the CON1 control group consists of students who would have been retained had the test policy been in existence one year earlier. Note that in Year 3 for each cohort the retained group is scoring substantially higher than the promoted group. This is the year of focus for the Greene and Winters (2006) study . It is easy to see how these researchers would have been led to believe that the test-based retention policy is working well.



Note that the gains in the retained group began to level off in Year 3 and, by year 5, the promoted Level 1 scoring students in the control group caught up with, if not surpassed, the retained students in proficiency. The initial gains of the retained group seem to have dissipated. This kind of initial gain and eventual falling back for retained students is typical of previous retention research (Holmes and Matthews 1984, Holmes 1989, Jimerson 2001).

A Different Set of Analysis Groups

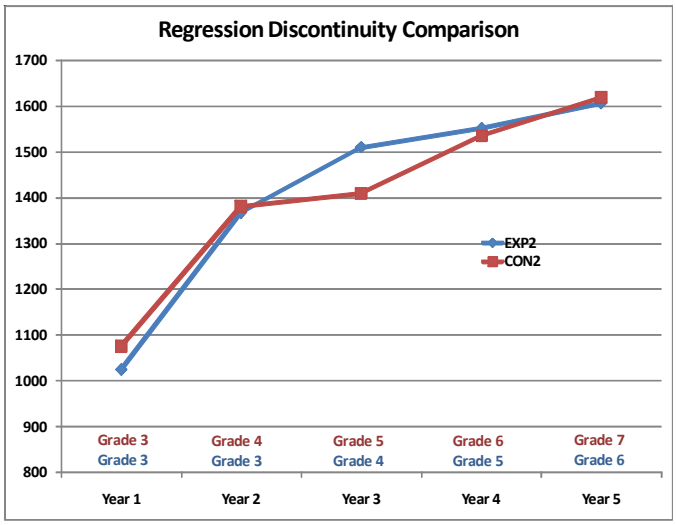
As was stated earlier, the perfect control group for the students retained in the first year of the FCAT requirements does not exist. The control group used in the across-year comparison was a close match -- students who would have been retained under the same FCAT requirements if the policy had been in existence one year earlier. One of the problems with this kind of control group is that there are other differences beside retention/non-retention between the two cohorts. As the two groups passed through the grade levels in different calendar years they may have experienced different school system influences. Changes in reforms such as vouchers, charter schools, or zone schools may have had differential effects on the two groups.

Another high-quality study of test-based retention was conducted in Chicago by Roderick and Nagaoka (2006). This study employed a different kind of research design, comparing students just below the cutoff score threshold, and thus retained, to students in the same grade cohort just above the cutoff score threshold, and thus promoted. This kind of research design was referred to as a “regression discontinuity” design.

Their results showed something quite different from the Greene and Winters (2006) study. They found that the retention policy in Chicago had a mild positive impact on the test performance of the retained students in the first year, but these gains disappeared or turned negative in the following year. In order to avoid the possible attribution of study result differences to differences in methodology, Greene and Winters (2006) also employed a regression discontinuity comparison in their study. Following the procedures in these studies, a similar kind of analysis was conducted in this study.

Regression Discontinuity Comparison

For this part of the analysis, a subset of the third-grade students who were retained in the first year of the test-based retention policy was identified. The experimental group 2 (EXP2) consisted of 975 students scoring within 50 points **below** the cutoff developmental scale score for retention, representing 3.9% of the entire 2002-03 third grade cohort. This experimental group of retained students was then compared to control group 2 (CON2) consisting of 1,274 students scoring within 50 points **above** the cutoff developmental scale score for retention, representing 5.1% of the entire 2002-03 third grade cohort.



The graph above shows the same kind of results as the across-year comparison. The EXP2 group of retained students just below the threshold were outperforming the CON2 group of promoted students just above the threshold during the third year of the study. Once again, Greene and Winters (2006), conducting their study in the Year 3, would naturally conclude that retention was working well. But, just as in the other comparison, the gains leveled off for the retained students until, in year 5, the two groups were essentially equivalent.

Discussion

This study has replicated the procedures of the Greene and Winters (2006) paper evaluating Florida's test-based promotion policy and has derived very different judgements. Where they concluded that the retention policy led to significant improvements in reading for the retained students, this study finds no ultimate advantages. However, it would be a mistake to interpret this study as some kind of indictment of the Greene and Winters work. Their interpretation was valid for the way the data looked after two years. The picture is quite different after four years. Even they say, in their conclusion section, that "We do not know whether the gains we have observed two years after students are retained will continue to hold, expand, or disappear over time." It appears now that the gains have essentially disappeared.

To say that these retained students are approximately where they would have been academically had they not been retained is to fail to notice another important characteristic of this cohort. They are also one grade level behind where they would have been and with classmates who, for the most part, are a year younger and were never retained. In the debate over the pro's and con's of retention, the practice before test-based decisions is commonly disparagingly referred to as an era of "social promotion." Perhaps we should be reminded that there is a social side to retention, as well.

We are also careful in our discussions to avoid referring to "passing" the FCAT or "flunking" third grade. But one wonders whether the third-graders, themselves, use these terms. It is impossible to fully assess the impact of being heldback on the self-esteem of the retained students and the academic expectations of their parents, classmates and teachers. One of the most common attributes of students who eventually drop out of school is having been retained sometime in their academic history. Over the years since the adoption of FCAT requirements for promotion from third grade, over 10,000 students in M-DCPS alone have been retained who would not have been retained otherwise. Many hundreds have been retained more than once in third grade. The idea behind third-grade retention is to give students who have substantial reading deficiencies the extra time and intensive instruction they may require to catch up. Undoubtedly, for some students this is just what was needed and, benevolently for them and society, may have set them back on a successful academic path. Equally undeniably, there are students who have not profited from retention enough to counterbalance what they have lost in the process. Given the enormous potential impact of this policy, it is incumbent upon us to continue reevaluating the consequences and rethinking our commitment.

References

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

Greene, J.P., Winters, M.A. (2006). “Getting further ahead by staying behind: a second-year evaluation of Florida’s policy to end social promotion.” Manhattan Institute for Policy Research, Civic Report No. 49.

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

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

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

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

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

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

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

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