AN OVERVIEW OF RESEARCH ON THE EFFECTIVENESS OF RETENTION ON STUDENT ACHIEVEMENT FOR NEW YORK CITY SCHOOLCHILDREN

A review of the research on the efficacy of retention-- the blanket strategy of using a single test to determine if a student should be held back-- demonstrates that retention impedes the educational progress of children, and leads primarily to lower achievement and higher drop out rates. Indeed, there are few issues in which there is such an overwhelming consensus among professionals as the negative effects of retention.

In March 2004 Mayor Bloomberg went ahead and forced through a policy of retention based on a single measure of performance for third graders. This was done despite widespread protest and the fact that a majority of members of the Panel on Education Policy opposed it. The policy only passed because three of the Panel members were removed hours before the vote.

The controversy surrounding this decision had its source in the abundant research of over a quarter of a century, clearly demonstrating that blanket retention policies hurt rather than help students and lead to much higher dropout rates. There is copious evidence that not only are such policies detrimental, but that in fact retention is even more harmful to children the higher their grade level, as discussed below.

The Folly of the Use of a Single Test to Determine Holdover

One of the inherent problems with the third grade policy and proposed fifth grade policy is the use of a single test. Test makers and educational researchers have concluded that a single test should not be the basis for significant educational decisions. The American Educational Research Association (AERA), the nation's largest professional

organization devoted to the scientific study of education, opposes their use in this way,[1] as does the National Board on Educational Testing, the International Reading Association,[2] and the National Council of Teachers of Mathematics, which argues that "far-reaching and critical educational decisions should be made only on the basis of multiple measures."[3] The Standards for Educational and Psychological Testing, developed by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, contain the following statement:

> "Any decision about a student's continued education, such as retention, tracking, or graduation, should not be based on the results of a single test, but should include other relevant and valid information."[4]

The National Academy of Sciences published a comprehensive report explaining in detail why the use of high-stakes testing is intellectually indefensible as well as counterproductive. [5] As the authors point out, "A student's score can be expected to vary across different versions of a test…. as a function of the particular sample of questions

---

[1] American Educational Research Association (AERA), Position Statement Concerning High-Stakes Testing in PreK-12 Education, Adopted July 2000; http://www.aera.net/about/policy/stakes.htm .

[2] International Reading Association, Summary of their position statement from *High-Stakes Assessments in Reading* (August 1999); http://www.reading.org/positions/high_stakes.html .

[3] The National Council of Teachers of Mathematics, "Position statement on High-Stakes Testing;" http://www.nctm.org/about/position_statements/highstakes.htm .

[4] AERA, op.cit., 2000; http://www.aera.net/about/policy/stakes.htm.

[5] Jay P. Heubert and Robert M. Hauser, eds., Committee on Appropriate Test Use, National Research Council, High Stakes: Testing for Tracking, Promotion, and Graduation, National Academy Press, 1999; http://www.nap.edu/catalog/6336.html.

asked and/or transitory factors, such as the student's health on the day of the test. Thus, no single test score can be considered a definitive measure of a student's knowledge."[6]

Harcourt and CTB McGraw Hill, the two largest companies that produce standardized tests, and the developers of New York City's 3rd grade reading and math exams, are on record opposing the use of their tests as the exclusive criterion for decisions about retention, because they can never be a reliable and/or complete measure of what students may or may not know. As Harcourt, the company that produces New York City's 3rd grade reading exam, has written:

"Another misuse of standardized achievement test scores is making promotion and retention decisions for individual students solely on the basis of these scores....Achievement test scores may certainly enter into a promotion or retention decision. However, they should be just one of the many factors considered and probably should receive less weight than factors such as teacher observation, day-to-day classroom performance, maturity level, and attitude."[7]

CTB-McGraw has the following statement on its website:

"No single test can ascertain whether all educational goals are being met. A variety of tests--or, multiple measures--is necessary to provide educators with a well-rounded view of what students know and can do. Just as different tests provide different information, no one kind of test can tell us all we need to know about a student's learning." [8]

In addition, as with all standardized tests, a substantial margin of error exists, inescapable given the nature of these exams. Thus, a number of students are likely to fail who would pass if the statistical uncertainties involved were taken into account.[9] There is

---

[6] Jay P. Heubert and Robert M. Hauser, op.cit., p. 3.

[7] Harcourt Brace Educational Measurement, Stanford Achievement Test Series, Ninth Edition: Guide for Organizational Planning, 1997, pp. 43-44.

[8] CTB McGraw Hill, "Educational Assessment: Four Principles to Consider"; http://www.ctb.com/articles/article_information.jsp?CONTENT%3C%3Ecnt_id=35477&FOLDER%3C%3Efolder_id=62821&bmUID=1064947932339

[9] For example, an analysis showed that those students who really belonged at the 50th percentile of the widely-used Stanford 9 test would be expected to score within five points of that mark only about 30% of the time in math, and only 42% of the time in reading. David Rogosa, Stanford University, "How Accurate

also the distinct possibility that the tests themselves may be flawed, or are scored incorrectly, as has occurred in the recent past. In such cases, more students would be unfairly held back, and their futures put at unnecessary risk.

New York City is not immune to these potential errors. During the April 2004 3rd grade exam, which was the new sole measure of promotion, officials reported that 1,300 students had previously seen test questions because they were used on earlier administrations of the test. This seriously implicated that validity of those test scores.[10]

Learning From History: Retention Policies Based on a Single Test Do Not Work

A retention policy based on a single test has been tried before in New York City. Research data on this program demonstrated that it was a clear failure, and the program was ended, but not until after huge sums of money were expended. In 1981 Chancellor Macchiarola launched the Promotional Gates program. Ten years later it was eliminated as a failed program. Under Macchiarola's plan 25,000 fourth through seventh graders were held back the first year. Low scores on citywide reading exams were the basis for the retention decision; math scores were included later. "Gates" required the hiring of an additional 1,100 teachers for the newly created retention classes capped at 18 students, and a summer school program.

The plan brought with it a huge financial burden. The additional teachers hired alone cost between $40 and $70 million, yet student outcomes were negative. The average summer school student made no improvements on their test score performance.

Are the STAR National Percentile Rank Scores for Individual Students?–An Interpretive Guide,' August 1999; http://www-stat.stanford.edu/~rag/ed351/drrguide.pdf.

[10] David M. Herszenhorn. "Retest Is Option For 3rd Graders Who Got Peek", *The New York Times*, April 29, 2004 Thursday, Late Edition - Final, Section B; Column 5.

After two years, retained students still showed no significant improvements over low-achieving counterparts who were promoted. Moreover, long term follow-up showed that 40% of the students who were retained eventually dropped out, compared to 25% of those with similar test scores who had been promoted. According to Ernest House, one of the authors of an evaluation mandated by the Mayor's office, "the Promotional Gates Program had retained tens of thousands of students at huge dollar and human costs without benefits."[11]

In light of this failure, on September 11, 1991, the Board of Education adopted a Resolution eliminating the Gates program. Recognizing the failure of the philosophy "Promotional Gates," the Board discontinued the program because "it did not sufficiently improve the achievement levels of participating grade 4 and 7 students."[12] The 1991 Resolution conceded that the Promotional Gates program had made them more likely to drop out:

> It has been determined that Promotional Gates had little positive impact on students. Each year more than one-third of the students who were held over and attended Gates classes still failed to meet promotional standards. Fourth grade holdovers were no more likely, three years later, to meet the seventh grade promotional standard than students who were promoted. A longitudinal study indicated that a disproportionate percentage of students held over in Gates classes became dropouts. National research has also confirmed the negative impact of retention policies by indicating that retention doubles the chances that a student will ultimately drop out. *There is no evidence, therefore, that holdovers make academic progress, although there is evidence that holdovers*

[11] E.R. House, R. Linn, R. and J. Raths, An Audit of the Evaluation of New York City's Promotional Gates Program. Four reports, October 1981, February 1982, April 1982, October 1982. Ernest House, one of the coauthors of the study, has summarized the report's conclusions in "The Predictable Failure of Chicago's Student Retention Program, November 1998; http://www.designsforchange.org/pdfs/houseChicago.pdf **.** See also the negative evaluation of the Gates program, carried out by R. Gampert, and Opperman, 1988, "Longitudinal Study of the 1982–83 Promotional Gates Students," cited in Hauser and Heubert, op.cit., footnote 13, p. 128.

[12] Board of Education 1991 Resolution at p. 1.

demonstrate greater social and emotional difficulties. (emphasis added). [13]

The dismal results of the "Gates" program have unfortunately been replicated in the single test policy enacted in the Chicago public schools. In 1996, Chicago instituted a policy of promoting students on the basis of their performance on the Iowa Tests of Basic Skills, a plan that cost approximately $100 million per year when it began.[14] Thousands of third, sixth, and eighth graders were retained.

A recent report by the Consortium on Chicago School Research tracked the effects of the policy on these students . Their findings reinforce the overwhelming professional consensus that single-test promotional standards do not work, and instead have deleterious effects. In their study, the Consortium utilized three different methodologies for the most accurate results. Each methodology compared the reading growth ability between low-achieving students just-above and just-below the promotional cutoff. The researchers found that in the third grade, the promotion policy had no effect on student performance.

Sixth graders experienced much worse outcomes. In the first post-gate year, retained sixth grade students achieved only a .82 growth in their reading abilities, while similarly low- achieving students who were promoted exhibited a growth of 1.19, a significant .37 difference. Two-years after the gate, again, there was an even greater

---

[13] Id. Also see Berger, J., "Fernandez to End a Policy on Holding Pulis Back," NY Times, Sec. A., p. 1 (Aug. 3, 1990).

[14] E.R. House, "The Predictable Failure of Chicago's Student Retention Program," November 1998, p. 17.

disparity of growth, a .44 difference. Retained sixth graders performed almost 25% worse than low-achieving sixth grade students who were promoted.[15]

Retention Policies Disproportionately Affect Black and Hispanic Students and Increases Dropout Rates

The large-scale retention policy now under consideration is not only counterproductive and extremely expensive, it is also inherently inequitable. The practice of retaining large numbers of New York City students on the basis of test scores alone is likely to disproportionately affect those who are poor and minority. Moreover, the policy of using high-stakes tests to make retention decisions has been shown to be more commonly used in school districts with high percentages of black and Hispanic students compared to the rest of the nation. Given the fact that research shows that these policies on balance are harmful to students who are subjected to them, their use appears to exacerbate rather than ameliorate racial and class differences.

To give some concrete examples in New York City; District 12, with a minority population of over 93% had only 19% of students achieve a level 3 or 4 on the 2004 administration of the math exam. Conversely, District 2 in Manhattan, which has a 70% white population, had over 60% of their students score 3 or 4 on the 2004 Math exam. Figures such as these all suggest that more minority than white students will be affected by promotional policies.

In addition to exacerbating racial and class differences, retention also leads to increased dropout rates, as was seen during the Gates program. A review of numerous

[15] Melissa Roderick and Jenny Nagoaka. "Ending Social Promotion: The Effects of Retention," *Charting Reform in Chicago Series*. The Consortium on Chicago School Research, March 2004, p.41.

studies on retention concluded that students who were held back were much more likely to eventually drop out. One study found that a student retained once was 40% to 50% more likely to dropout of school, and 90% more likely if retained twice.[16] A longitudinal study of more than 12,000 students published in the *American Journal of Education* found that, after controlling for student background and academic achievement, being held back before the 8th grade increased the likelihood of dropping out by the 12th grade by more than 200%. Furthermore, "students who were held back before the 8th grade were more than four times as likely as students who were not held back to not complete high school or receive a GED" six years later.[17]

Conclusion

With such overwhelming evidence as to the negative effects of retention it is puzzling from an educational policy perspective why this administration wants to implement a program of retention for third and fifth graders. The program has failed to help New York City students in past, and has failed to assist students in school districts across the country for the last twenty-five years.

It is clear that the single test retention policy is not supported by the testing companies or by the research, and disproportionately affects black and Hispanic students. The funds spent on retaining students would be better used to provide smaller classes, increase access to pre-kindergarten, and offer intensive intervention for students who

---

[16] Mann, D. (1987). Can we help dropouts? Thinking about the undoable. In G. Natriello (Ed.), School dropouts: Patterns and policies (pp. 3–19). New York, NY: Teachers College Press. In Shane R. Jimerson, Gabrielle E. Anderson, and Angela D. Whipple. "Winning the Battle and Losing the War: Examining the relation between grade retention and dropping out of high school," *Psychology in the Schools*, Vol. 39(4), 2002, p.12.

[17] W. Rumberger and K.A. Larson, Student Mobility and the Increased Risk of High School Dropout, *American Journal of Education*, November 1998.

have fallen, or are at-risk of falling behind, programs that have been demonstrated to improve student performance. Instead, the Mayor is poised to repeat a dismal history of failure in this area.