High-Stakes Testing and Student Achievement: Are High-Stakes Tests Helping or Hurting America's Children?

MAPP Discussion Paper #3

Jennifer Kleindienst
Joseph Rinaldo

Hamilton College

April 2007

Topic Definition

No Child Left Behind is a piece of very controversial and complex legislation. Much of the controversy surrounds the effectiveness of this act, and we wanted to find out if No Child Left Behind is affecting student achievement positively, neutrally, or negatively. This question proved too complex because there are so many components of No Child Left Behind, so we decided to focus on the effects of high-stakes testing, an integral part of No Child Left Behind, on student achievement. According to the Education Policy and Leadership Center, high-stakes testing is the "use of test scores to make decisions that have important consequences for individuals," including graduation exams, grade promotion exams, exams that determine school, teacher, or student "access to resources and special opportunities", or exams that measure teacher quality (Pearlman, 2001). High-stakes tests have proved so controversial precisely because of the consequences that are associated with high-stakes tests. By reviewing relevant literature, we seek to discover the effects of high-stakes testing on America's schools.

Policy Environment

American public education has relied on exams to test student achievement for nearly a century. Until the post-war era, tests were usually given by individual teachers, schools, or school districts, but in the 1950s, Cold War fears began the push for the standardization of education with the intention of improving student achievement. By the 1970s, statewide standardized tests, often known as "minimum competency tests" were widely in use to measure students' basic skills. In 1983, the National Commission on Education under President Ronald Reagan released *A Nation at Risk*, a report that called for a high-stakes testing movement to respond to growing concerns that basic skills tests were promoting low standards and low student achievement. Although this report was later found to be historically inaccurate, it has been very effective in convincing states and policymakers that accountability is the key to improving student achievement (Amrein & Berliner, 2002).

The conclusions of *A Nation at Risk* helped to fuel the 2001 No Child Left Behind Act, a reauthorization of the 1965 Elementary and Secondary Education Act. A key component of No Child Left Behind is a high-stakes testing program through which the Act aims to bring the nation's children to a common educational standard by 2014. One particular goal of No Child Left Behind is to close the testing achievement gap between white and minority students by holding students, teachers, and schools accountable for poor test scores and imposing sanctions on underperforming schools (Cronin, 2005).

The current high-stakes testing design is modeled after the Texas Assessment of Academic Skills (TAAS) exam, a test used in Texas to measure student achievement from 1992 to 2004 under the governorship of George W. Bush. As part of the program, high-performing schools are labeled as "exemplary" and receive money to spend on pet projects. Schools that do poorly are labeled as "inadequate" and are taken over by the state or a private organization if they do not improve by the following year. School labels are widely publicized, which increases the pressure on students, teachers, and schools to do well (Carnoy et al., 2001). After the implementation of TAAS, test scores improved dramatically, causing many politicians to label the TAAS the "Texas Miracle". What most lawmakers failed to notice, however, was that while TAAS scores

increased, Texas students' performance on the National Assessment of Educational Progress (NAEP), a non-high-stakes standardized test, did not increase. Nevertheless, upon ascending to the presidency, President Bush transformed the TAAS program from the state to the federal level.

Under No Child Left Behind, high-stakes tests vary between states; there is no unifying federal exam. Because so much depends on the achievement scores of the students, some states have designed tests that are easier for students, or teach the students the test material, which makes it appear as though student achievement on the tests is improving. This allows schools to seem as though they are improving academically and creates the illusion that state high-stakes tests are helping to improve achievement ("Reality…").

Political Environment

When No Child Left Behind was first introduced, it had large bipartisan support, as legislators were looking for a way to improve American student achievement and close the achievement gap. High-stakes testing is especially attractive because it sets standards, punishes schools that do not meet these standards, and rewards schools that do meet the standards. Supporters of testing, who are frequently Republicans, claim that the accountability of the tests will improve academic achievement and close the achievement gap between white and minority students, pointing to the Texas Assessment of Academic Skills exam. When No Child Left Behind was first introduced, opponents were primarily Democrats, but a sizable number of Republicans are now opposed to high-stakes testing, expressing fears that high-stakes testing will not improve academic achievement and may disproportionately affect minority students (Cronin, 2005). As noted earlier, however, the TAAS exam has undergone intense scrutiny recently because of allegations that test scores increased because of "teaching to the test" ("Reality…"). In adapting the TAAS for the entire nation, it appears that lawmakers did not carefully assess the test's effects on alternative measures of student achievement.

Theory

At first glance, high-stakes testing is very attractive because, in theory, accountability will pressure students to succeed, and it is understandable why policymakers would want to implement such tests. High-stakes testing has become popular because of two assumptions: greater rewards will make students try harder and the meaning of rewards and punishments is the same for all students (Madaus and Clarke, 2001). From our research, it appears as though these assumptions are not true for all students, as accountability is not improving student achievement.

Methodology

We began our search for relevant studies by searching several databases for information on the relationship between high-stakes testing and student achievement. As we originally planned to do a meta-analysis, we were looking for quantitative studies that utilized high-level statistical procedures. In our preliminary search, we found no studies when we searched the Humanities & Social Science Index Retrospective database, EconLit, the National Bureau of Economic Research (NBER), Social Sciences Abstracts, and Sociological Abstracts databases with the

search terms “high stakes testing” or “standardized testing” or “standards-based testing” AND “achievement” or “effectiveness” or “success” AND “quantitative”. Using the same search terms, we found twelve articles in the Educational Resource Information Center (ERIC) database and one article in the Professional Development Collection (PDC) database, but it was a duplicate of an article that turned up in the ERIC search. We also searched the Dissertation Abstracts database with the same search terms, finding fifteen relevant articles. Thus, our preliminary search yielded twenty-seven articles in our search of eight databases.

To reduce the effects of publication bias, we contacted several college and university professors who specialized in educational research and whose names we found through our database searches. The researchers were BetsAnn Smith of Michigan State University, David Herrington of Texas A&M University at Prairie View, Jay Heubert of Columbia University Teachers College, and Ruth Knudson of California State University at Long Beach. We explained that we were doing a meta-analysis and that we were looking for additional quantitative research relevant to the question of the effects of high-stakes testing on student achievement. We received a reply from David Herrington, but he informed us that he could not supply us with additional studies.

As we began to read the articles that had turned up in our database search, we quickly eliminated all but three of the studies. Most of the studies were surveys, case studies, or qualitative research; thus, they would not aid our meta-analysis. To expand our list of literature, we searched the bibliographies of studies from our original master list to find more relevant documents.

When we began to code the studies for our meta-analysis, we encountered great difficulty. Although we had a master list with seventeen articles, we found that, at second glance, many of these studies did not examine the relationship between high-stakes tests and student achievement. The scarcity of high-level statistical studies compounded the problem, and we were unable to figure out how to code many of the comparison or correlation analyses, so we contacted Audrey Amrein-Beardsley and David Berliner, two educational researchers at Arizona State University who have researched this topic extensively. Dr. Amrein-Beardsley informed us that the type of research that would yield effect sizes is unavailable. She continued, saying “there is never a clear indication of states with and without high-stakes testing policies and the degrees to which stakes matter vary across states[,] making these types of dichotomous analyses difficult and often indefensible” (Amrein-Beardsley, 2007). Because of this response and the impossibility of coding most of our studies, we determined that a meta-analysis would not be possible and chose instead to do a literature review of the available research.

Results

Although we found very little high-level statistical research that examined the effects of high-stakes testing on student achievement, the studies we did find can be divided into four main methodological categories. The bulk of the studies compared alternative test scores in high-stakes states to those in states without high-stakes tests, and used either a cross-sectional or cohort analysis. Other studies correlated alternative test scores in high-stakes and non-high-stakes states examining a cross-section of the population or a cohort group. We also found one study that was a cross-sectional experiment, and another that was a cross-sectional regression

analysis. The most common alternative tests used in these analyses were the National Assessment of Educational Progress (NAEP) Exam, Scholastic Achievement Test (SAT), Advanced Placement (AP) Exams, and American College Test (ACT).

Although we found a wide variety of results, it is important to note that results at higher statistical levels, like those results from the regression, carry more weight because information from the top of the statistical hierarchy is more meaningful than information from the bottom. Higher-level statistical procedures are more reliable because they utilize methods to control for confounding factors. For example, a regressional study that analyzed the effects of high-stakes testing on student achievement, measured by alternative test scores, between states with and without high-stakes tests and controlled for confounding factors, like race and socioeconomic background, would be more dependable than a study that used simple comparison analysis to examine the differences in alternative test scores between states with high-stakes tests and states without such tests. Out of nine studies, some of which used more than one methodology, five studies utilized comparison and cross-sectional analysis, one study used comparison and cohort analysis, two studies used a cross-sectional correlation, two used cohort correlation analysis, one used an experiment on a group of children, and one used a cross-sectional regression analysis. We further divided the results to show effects on math, reading, and other (overall or other subject) tests. The results are listed in the three tables below and in the appendix.

Methodology		**Reading Exam Effects**	
		4th grade	8th grade
Simple Comparison	Cross-sectional Analysis	Rosenshine – small positive effect Amrein & Berliner – inconsistent effect Amrein-Beardsley & Berliner – no significant effect	
	Cohort Analysis	Amrein & Berliner – very small positive effect (but NAEP gain may not be because of high-stakes test)	
Correlation Analysis	Cross-sectional Analysis	Nichols et al. – very small positive effect (close to no effect)	Nichols et al. – very small positive effect (close to no effect)
	Cohort Analysis	Nichols et al. – very small positive effect (close to no effect)	Nichols et al. – very small positive effect (close to no effect)
Regression Analysis	Cross-sectional Analysis	Cronin – very small negative effect (close to no effect)	Cronin – very small negative effect (close to no effect)

Methodology		Math Exam Effects		
		4th grade	8th grade	Graduation Exams
Simple Comparison	Cross-sectional Analysis	Rosenshine – small positive effect Amrein & Berliner – small negative effect (exclusion rate effect) Amrein-Beardsley & Berliner – large positive effect (exclusion rate effect) Madaus & Clarke – large negative effect Neill & Gayler – no effect	Rosenshine – small positive effect Amrein & Berliner – small negative effect (exclusion rate effect) Amrein-Beardsley & Berliner – no significant effect Madaus & Clarke – large negative effect Neill & Gayler – large negative effect	Madaus & Clarke – large negative effect (NAEP, ACT, ACT)
	Cohort Analysis	Amrein & Berliner – large negative effect (exclusion rate effect)		
Correlation Analysis	Cross-sectional Analysis	Nichols et al. – very small positive effect (close to no effect) Braun – very small positive effect (close to no effect)	Nichols et al. – very small positive effect (close to no effect) Braun – very small positive effect (close to no effect)	
	Cohort Analysis	Nichols et al. – very small positive effect (close to no effect) Braun – very small negative effect (close to no effect) (1992/1996), very small positive effect (close to no effect) (1996/2000)	Nichols et al. – very small positive effect (close to no effect) Braun – very small negative effect (close to no effect) (1992/1996), very small positive effect (close to no effect) (1996/2000)	
Regression Analysis	Cross-sectional Analysis	Cronin – very small negative effect (close to no effect)	Cronin – very small negative effect (close to no effect)	

Methodology		Other Exam Effects	
		8th grade	Graduation Exams
Simple Comparison	Cross-sectional Analysis		Amrein & Berliner – large negative effect (ACT, SAT), no effect (AP)
Experiment	Cross-sectional Analysis	Berube – large negative effect (Virginia SOL Science Exam)	

Simple Comparison Studies

Simple comparison studies utilize a very simple methodology, comparing mean alternative test scores between states with and without high-stakes tests. Some studies look at a cross-section of the population, examining test scores in a particular grade over time, while other studies examine a cohort of students, following a group of students over time. The studies that used cross-sectional simple comparison found a wide mix of results – some studies found large negative effects, some found large positive effects, some found no effect, and some found effects in the middle. If our study included only these articles, our conclusions would be murky at best. Barak Rosenshine compared fourth grade math, eighth grade math, and fourth grade reading National Assessment of Educational Progress (NAEP) score increases in high-stakes and non-high-stakes states. On average, all states showed a score increase on the NAEP, but the increases were 1.05 to 2.23 points higher in high-stakes states. Rosenshine concludes that although there were apparent increases, "attaching accountability to statewide tests…was not an effective policy in all states" (Rosenshine, 2003).

Audrey Amrein and David Berliner compared a larger group of exams, including the fourth and eighth grade NAEP exams, American College Test (ACT), Scholastic Achievement Test (SAT), and Advanced Placement (AP) exams, finding mixed results using both cross-sectional and cohort analysis. There were some gains in achievement at the lower grade levels, but the high school level tests indicated that the presence of high-stakes graduation exams decreased SAT and ACT performance and had no effect on AP performance. In addition, Amrein and Berliner attribute a large portion of test score gains to the exclusion effect, which occurs when states and schools intentionally or unintentionally exclude students from taking high-stakes tests because these students are underperforming and would bring down the school's scores, putting the school at risk of state takeover. Amrein and Berliner note that in "states with high-stakes tests, between [zero and forty-nine percent] of the gains in NAEP scores can be explained by increases in rates of exclusion". This indicates that high-stakes exams are not having the desired effect (Amrein and Berliner, 2002). In their 2003 study, Amrein-Beardsley and Berliner find that high-stakes tests have an insignificant effect on student achievement on the fourth grade reading and eighth grade math exams. While they observed a large positive effect of high-stakes testing on fourth grade math NAEP scores, they also attribute this to the exclusion effect (Amrein-Beardsley and Berliner, 2003).

Researchers Monty Neill and Keith Gayler compared raw scores on the fourth and eighth grade NAEP in states with and without high-stakes tests. Their findings were interesting – at the fourth grade level, mean score differences between high- and low-stakes state scores and the percent of students scoring at the "proficient" level or above were not statistically significant, and the percent of students in high-stakes states scoring at the "basic" level or above was significantly higher than in low-stakes states. By eighth grade, however, all scores – mean scores, the percent of students scoring at "basic" or above, the percent of students scoring at "proficient" or above, and the percent of students scoring at the "advanced" level – showed statistically significant differences that favored the low-stakes states (Neill and Gayler, 2001). This indicates that high-stakes testing is not working to improve overall student achievement, but it is important to note that Neill and Gayler do not control for confounding factors.

In a study of differences in alternative testing achievement based on race, George Madaus and Marguerite Clarke compared white, black, and Hispanic math scores on the fourth, eighth, and graduation-level NAEP exams, noting that the achievement gap did not significantly decrease between 1973 and 1996 on any of the tests. Black scores increased gradually on all three tests between 1973 and 1986, and steeply between 1986 and 1990 on the graduation-level exam. However, black student math scores leveled off from 1986 to 1996 at the fourth grade level and have fallen on the eighth grade and graduation exams. Hispanic scores on the fourth grade math exam fell from 1973 until 1994, when they began to rise, while eighth grade and graduation-level exams saw no significant change in this period. Black and Hispanic achievement has also not improved significantly on the SAT and ACT math exams (Madaus and Clarke, 2001). These studies indicate that high-stakes testing is not helping to improve non-Asian minority achievement and is not helping to close the achievement gap.

All researchers who performed comparison analyses agreed that more research needed to be done on the effects of high-stakes testing. Rosenshine advocates for more research so that policymakers can understand why these increases occurred, commenting that it is "inappropriate to simply use these results…and blindly require all states to impose consequences" (Rosenshine, 2003). Amrein and Berliner agreed, saying "high-stakes tests being used today do not, as a general rule, appear valid as indicators of genuine learning, of the types of learning that approach the American ideal of what an educated person knows and can do" (Amrein and Berliner, 2002).

Correlation Studies

Two studies used correlation analysis to measure the effects of high-stakes testing pressure on students. The first study, by Sharon Nichols, Gene Glass, and David Berliner, correlated fourth and eighth grade performance on the NAEP in states with and without high-stakes tests, breaking the results down by race. They concluded that high-stakes testing had a very weak positive effect on student achievement, as correlations ranged from -0.27 to 0.30, with all but two correlations indicating a positive relationship. Although this study controlled for confounding factors to a certain extent by breaking down scores into white, black, and Hispanic subgroups, the authors acknowledge the existence of additional confounding factors (Nichols et al., 2006). The second study by Henry Braun observed correlations in NAEP math scores. In his cross-sectional analysis, Braun found small positive achievement increases in states with high-states tests, but when he performed a cohort analysis, students in high-stakes states did not perform as well as students in low-stakes states (Braun, 2004). Both Braun and Nichols et al. found weak correlations between student achievement and the presence of high-stakes tests, indicating that overall, high-stakes tests have no strong effect on student achievement. Nichols et al. believe that because there is "no dependable or compelling evidence that the pressure associated with high-stakes testing leads to increased achievement, there is no reason to continue the practice of high-stakes testing" (Nichols et al., 2006). At the very least, there should be much more research on this subject.

Experimental Studies

Clair Berube, a former eighth-grade science teacher who is now an education professor at Wagner College on Staten Island, performed an experiment on his students. All middle school

science students were required to take the Virginia Standards of Learning (SOL) test, which used only multiple choice questions to test the students' knowledge. A week after the students had taken the SOL, Berube gave each eighth grader who had passed the state exam an additional "comprehension measurement" test he designed in which students had to defend their answers on the state exam to demonstrate their knowledge of the test material. Berube found that 71.25 percent of the eighty eighth graders in six classrooms who had passed the SOL could not pass his version of the test. This is a likely indication that many of the eighth grade science teachers were "teaching to the test"; however, Berube's conclusion must be discounted because his sample size was only eighty students. Berube explained the results he found, stating, "tests are seen as symbols of order, control, and attainment. However, if important decisions—jobs and governmental funding—rely on the outcome of high-stakes tests, then teachers will only teach to the test, allowing test content to define the curriculum" (Berube, 2004).

Regression Studies

In the last study, John Cronin, G. Gage Kingsbury, Martha McCall, and Branin Bowe performed a cross-sectional regression analysis on students in third through eighth grade to measure the effects of the testing component of No Child Left Behind. Cronin et al. found that while high-stakes test *scores* increased from the 2001-2002 school year to the 2003-2004 school year, student *achievement*, as measured by test scores on the NAEP, fell in this period. It is important to note that the changes in student achievement growth were very small, with an effect size of -0.04 for third through eighth math scores and a -0.02 growth decrease for reading scores; thus the study indicates that high-stakes tests have virtually no effect on real student achievement (Cronin, 2005).

Conclusions

Although educators and lawmakers may not agree on the effects of high-stakes testing, one thing is certain: there is insufficient high-level statistical research on high-stakes testing and the tests' effects on student achievement. It is hard to believe that high-stakes tests could be instituted on a large scale, forming the basis for the United States' entire educational system, when little high-quality research has been done to evaluate whether the tests are worthwhile indicators of student achievement and if they are helping or harming students. What little research is available does not support the argument for high-stakes testing. Nearly every available study included a mention of uncertainty because of a lack of enough information. Most available studies on the relationship between high-stakes testing and student achievement are based on clinical judgments, surveys, or comparisons of raw figures. We found only one regression that explained the relationship between high-stakes testing and student achievement, which is indicative of the dearth of available research. More research is especially needed to examine the effect of high-stakes testing on high-school age students.

Controlling for confounding factors is an important component of research, and studies that do more to control for confounding factors are more reliable than studies that do less to control for these factors. Some studies did better than others at controlling for confounding factors, often performing cohort analyses or examining testing's effects on a subgroup of the population. We found that, in general, studies that utilized simple comparison or other lower-level statistical

research had fewer controls for moderator variables than studies that used higher-level statistical research.

Of the five studies that used simple comparison, only three controlled for confounding factors. Amrein and Berliner compared alternative test scores and adjust for changes in alternative test scores by examining average scores before and after the implementation of high school graduation exit exams. They also controlled for changes in the percentages of excluded students. With these controls, they found a mixture of positive, negative, and inconsistent effects (Amrein and Berliner, 2002). Amrein-Beardsley and Berliner's study controlled for the exclusion rate by breaking down NAEP score differences into states with clear gains or losses and states with unclear gains or losses. A state displays a clear gain or loss when "a state's scores increase [or decrease] while the rates by which students are exempted from the NAEP stay the same or decrease", while unclear gains or losses occur when "a state's scores increase [or decrease] while the rates by which students are exempted from the NAEP increase". Using these controls, Amrein-Beardsley and Berliner found mainly insignificant effects because of the exclusion rate effect (Amrein-Beardsley and Berliner, 2003). Madaus and Clarke examined test score differences between ethnic groups to control for racial moderator variables, and found negative effects in all cases (Madaus and Clarke, 2001). Neither Rosenshine nor Neill and Gayler controlled for any confounding factors. While Neill and Gayler found a mix of negative and no effects, Rosenshine was one of the few researchers who found positive effects of high-stakes testing on student achievement (Rosenshine, 2003; Neill and Gayler, 2001).

In his correlation analysis, Braun controlled for changes in NAEP scores by looking at exclusion rate changes from year to year. He also correlated changes in test scores to each state's "policy score", which was calculated by grading "each state on each of 22 policy activities organized into four categories: content standards, performance standards, aligned assessments, and professional standards. Grades were assigned on a three point scale: does not have such a policy (0), is developing one (1), or has enacted such a policy as of 1996 (2)". Both of these controls were performed using both cross-sectional and cohort analysis (Braun, 2004). Nichols et al. controlled for changes in the exclusion rate and for differences in Hispanic, African-American, and white test scores.

Berube's experiment controlled for differences in teaching styles – constructivist teachers place an emphasis on problem-based learning and student discovery, while traditional teachers tend to rely more on lecture methods to educate students. Berube hypothesized that students from constructivist classrooms would perform better than students in traditional or mixed constructivist/traditional classrooms, but his results indicated that students from constructivist classrooms actually performed worse than students in mixed classrooms (Berube, 2004).

Cronin et al.'s regression controlled for race and ethnicity, and measured student growth by comparing differences in beginning and end student test assessment scores. By using multivariate regression analysis, they were able to examine relationships between significant dependent and independent variables and could calculate standard deviations and effect sizes to standardize results. Using ANOVA (Analysis of Variance between groups), Cronin et al. determined that the ethnicity of students was the variable with the most influence on student achievement and that the presence of state testing was the second most influential variable on

alternative test scores (Cronin, 2005). Such controls make results more reliable and are more informative.

Despite these enormous obstacles, we were still able to draw several conclusions from our research. We conclude that overall, high-stakes testing has virtually no effect on student achievement, although we note that it is difficult to draw strong conclusions because available high-level statistical research is scarce. There were a few cases in which math scores appeared to have positive effects on student achievement using cross-sectional analysis, but once a cohort analysis was performed on the same testing data, the positive effect vanished (Braun, 2004; Amrein and Berliner, 2002). The few studies that found results in favor of high-stakes testing noted that these positive achievement effects were small, and a substantial number of studies attributed apparent achievement increases to the exclusion of students from test-taking, also known as the exclusion rate (Amrein and Berliner, 2002; Amrein-Beardsley and Berliner, 2003). In addition, the studies that indicated high-stakes tests have a significant positive effect on student achievement all utilized low-quality statistical methodologies, which discounts their findings. In the ten states with the highest dropout rates, nine used graduation exit exams, whereas five of the ten states with the lowest dropout rates did not have testing programs, which may be indicative of the negative effects of graduation exams (Madaus and Clarke, 2001), however researchers Martin Carnoy, Suzanna Loeb, and Tiffany L. Smith used regression analysis and found a positive relationship between lower dropout rates and higher scores on the TAAS exam (Carnoy et al., 2001). More research must be done to examine the relationships between high-stakes testing and alternative measures of student achievement, like the dropout rate, retention rate, and graduation rate before solid conclusions can be drawn. Ultimately, we discovered no overwhelming evidence in favor of high-stakes testing and believe that such testing practices should not continue unless drastic changes are made to increase the breadth of such tests.

One unintended side effect of high-stakes testing is its possible negative effect on poor and minority students, but because of limited research, we cannot draw any definite conclusions. One of the goals of the No Child Left Behind legislation was to close the achievement gap, but the available research indicates that high-stakes tests do at best nothing to close the gap, and may even increase the gap (Madaus and Clarke, 2001). In a United States Department of Education study on National Assessment of Educational Progress (NAEP) scores from 1990 to 2005, the achievement gap decreased in reading, but average reading scores decreased in the same period, so no true gains were made. The achievement gap in math did not decrease during this time, which should indicate that high-stakes tests are not improving student achievement ("Achievement", 2005). Madaus and Clarke had similar findings in their studies of NAEP, SAT, and ACT math scores (Madaus and Clarke, 2001). Minorities have improved test scores in some cases, but the achievement gap between white and minority scores is still large.

There is growing discontent among American educators that high-stakes tests are not working to improve student achievement, which should be of major concern, as high-stakes testing is a major component of No Child Left Behind. If, as our conclusions suggest, high-stakes testing is not affecting students as policymakers intend, there is no reason to continue such testing practices. Placing so much pressure on the teachers to have their students do well on the tests is not the best way to improve the education of the children, as this pressure often forces teachers in

underperforming schools to "teach to the test", essentially transforming the classroom into a test preparation center. This greatly limits the breadth of education students in underperforming schools receive. In the process, many students are being deprived of a well-rounded education, which can encourage children to think creatively and prepare them for life beyond the classroom.

Policy Implications

High-stakes tests have enormous policy implications. Millions of dollars are currently being allocated for high-stakes testing under No Child Left Behind, with virtually no high-level statistical evidence that the tests are effective measures of student achievement. Thus, it is unclear to policymakers if No Child Left Behind should continue in its present form, be drastically amended, or eliminated. No Child Left Behind is currently up for reauthorization, and President Bush wants to allocate more money to continue this program, yet there is scarce high-level evidence on the effects of the testing component of No Child Left Behind. Because of a lack of statistical research on this topic, policymakers are forced to rely mainly on clinical predictions and low-level statistical research to base their assessment of No Child Left Behind. After a thorough analysis of the available literature, we conclude that standardized tests are necessary as a uniform measure of student achievement, but that stakes need to be eliminated, as the available studies have indicated that attaching stakes to standardized tests does not provide an achievement benefit for the students.

References

"Achievement Gap: Differences in White-Black and White-Hispanic 4th- and 8th-grade average reading and mathematics scores: Various years, 1990–2005." *U.S. Department of Education, National Center for Education Statistics*. December 2005. <http://nces.ed.gov/programs/coe/2006/charts/chart14.asp?popup=true>.

Amrein, Audrey and David Berliner. "High-Stakes Testing, Uncertainty, and Student Learning." *Education Policy Analysis Archives*, 10.18 (2002). Accessed via <http://epaa.asu.edu/epaa/v10n18/>.

Amrein-Beardsley, Audrey. "Re: Questions about Studies." E-mail to Jennifer Kleindienst. 3 April 2007.

Amrein-Beardsley, Audrey and David Berliner. "Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine." *Education Policy Analysis Archives*, 11.25 (2003, August). Accessed via <http://epaa.asu.edu/epaa/v11n25/>.

Berube, Clair T. "Are Standards Preventing Good Teaching?" *Clearing House*, 77.6 (July/August 2004): 264-267.

Braun, Henry. "Reconsidering the impact of high-stakes testing." *Educational Policy Analysis Archives*, 12.1 (2004): 1-40. Accessed via <http://epaa.asu.edu/epaa/v12n1/>.

Carnoy, Martin, Suzanna Loeb, and Tiffany L. Smith. "Do Higher State Test Scores in Texas Make for Better High School Outcomes?" Consortium for Policy Research in Education: University of Pennsylvania Graduate School of Education. November 2001.

Cronin, John, G. Gage Kingsbury, Martha S. McCall, and Branin Bowe. "The Impact of the No Child Left Behind Act on Student Achievement and Growth." *Northwest Evaluation Association Growth Research Database*. April 2005. Accessed via <http://www.nwea.org/assets/research/national/NCLBImpact_2005_Study.pdf>.

Gussner, William. "Schools Alone Can't Make Sure No Child Is Left Behind." *St. Louis Post-Dispatch (Missouri)*. 19 April 2001. B7.

Madaus, George and Marguerite Clarke. "The Adverse Impact of High-Stakes Testing on Minority Students: Evidence from One Hundred Years of Test Data." Ed. *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*. Orfield, Gary and Mindy Kornhaber. New York: The Century Foundation, 2001, 85-106.

Neill, Monty and Keith Gayler. "Do High-Stakes Graduation Tests Improve Learning Outcomes? Using State-Level NAEP Data to Evaluate the Effects of Mandatory

Graduation Tests." Ed. *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*. Orfield, Gary and Mindy Kornhaber. New York: The Century Foundation, 2001, 107-125.

Nichols, Sharon L., Gene V. Glass, and David C. Berliner. "High-stakes testing and student achievement: Does accountability pressure increase student learning?" *Education Policy Analysis Archives* 14.1 (2006). Accessed via <http://epaa.asu.edu/epaa/v14n1/v14n1.pdf>.

Pearlman, Mari. "High-stakes Testing: Perils and Opportunities." Pennsylvania Education Policy Forum. Harrisburg, PA. 4 April 2001.

"Reality Testing NCLB." *FairTest: The National Center for Fair & Open Testing*. Accessed via <http://www.fairtest.org/nattest/Reality_Testing_NCLB.html>.

Rosenshine, Barak. "High-Stakes Testing: Another Analysis." *Education Policy Analysis Archives*, 11.24 (2003). Accessed via <http://epaa.asu.edu/epaa/v11n24/>.