

BUILDING ON THE BASICS:
The Impact of
High-Stakes Testing on
Student Proficiency in Low-
Stakes Subjects

Marcus A. Winters, Ph.D.
Senior Fellow, Manhattan Institute

Jay P. Greene, Ph.D.
Senior Fellow, Manhattan Institute

Julie R. Trivitt, Ph.D.
*Assistant Professor,
Arkansas Tech University*

School systems across the nation have adopted policies that reward or sanction particular schools on the basis of their students' performance on standardized math and reading tests. One of the most frequently raised concerns regarding such "high-stakes testing" policies is that they oblige schools to focus on subjects for which they are held accountable but to neglect the rest. Many have worried that the limited focus of these policies could have an unintended negative effect on student proficiency in other subjects, such as science, that are important to the development of human capital and thus to future economic growth.

This paper uses a regression discontinuity design utilizing student-level data to evaluate the impact of sanctions under Florida's high-stakes testing policy on student proficiency in science. Under that state's A+ program, every public school receives a letter grade from A to F that is based primarily upon its students' performance on the state's standardized math and reading exams. Students in Florida were also administered a standardized exam in science, but this test was low-stakes because its results held no consequences under the A+ program or any other formal accountability policy.

Previous research has found that the rewards and sanctions of receiving an F grade in the prior year led to improved gains in student proficiency in the high-stakes subjects of math and reading. This current paper is the first to evaluate the impact of the incentives under this high-stakes testing system on student proficiency in science. This paper adds to a sparse previous literature quantitatively evaluating whether high-stakes testing policies have "crowded out" learning in a low-stakes subject.

The primary findings of the study are:

- The F-grade sanction produced after one year a gain in student science proficiency of about a 0.08 standard deviation. These gains are similar to those in reading and appear smaller than the gains in math that were due to the F sanction.
- There is some evidence to suggest that student science proficiency increased primarily because student learning in math and reading enabled that increase. That is, learning in math and reading appear to contribute to learning in science.

ABOUT THE AUTHORS

MARCUS A. WINTERS is a senior fellow at the Manhattan Institute. He has conducted studies of a variety of education policy issues including high-stakes testing, performance pay for teachers, and the effects of vouchers on the public school system. His research has been published in the journals *Education Finance and Policy*, *Teachers College Record*, and *Education Next*. His op-ed articles have appeared in numerous newspapers, including the *Wall Street Journal*, *Washington Post*, and *USA Today*. He received a B.A. in political science from Ohio University in 2002 and a Ph.D. in economics from the University of Arkansas in 2008.

JAY P. GREENE is endowed chair and head of the Department of Education Reform at the University of Arkansas and a senior fellow at the Manhattan Institute. He conducts research and writes about topics such as school choice, high school graduation rates, accountability, and special education.

Dr. Greene's research was cited four times in the U.S. Supreme Court's opinions in the landmark *Zelman v. Simmons-Harris* case on school vouchers. His articles have appeared in policy journals such as *The Public Interest*, *City Journal*, and *Education Next*; in academic journals such as *The Georgetown Public Policy Review*, *Education and Urban Society*, and *The British Journal of Political Science*; as well as in newspapers such as the *Wall Street Journal* and the *Washington Post*. He is the author of *Education Myths* (Rowman & Littlefield, 2005). Dr. Greene has been a professor of government at the University of Texas at Austin and the University of Houston. He received a B.A. in history from Tufts University in 1988 and a Ph.D. from the Department of Government at Harvard University in 1995.

JULIE R. TRIVITT is an assistant professor of economics at Arkansas Tech University. She earned her Ph.D. in economics from the University of Arkansas in December 2006. Her focus was health economics and applied econometrics. Since completing her Ph.D., she has worked as a senior research associate in the Department of Education Reform at the University of Arkansas and as a consultant on education projects to the World Trade Center in Brescia, Italy. Her research agenda focuses on human capital, health, labor, and education economics.

CONTENTS

- 1 Introduction
- 3 Florida's A+ Accountability Program
- 4 Data and Method
- 5 The Impact of the F-Grade Sanction on Student Proficiency in High- and Low-Stakes Subjects
- 6 Understanding the Effects of the F-Grade Sanction on Science Proficiency
- 7 Summary and Discussion
- 8 Endnotes
- 9 References

BUILDING ON THE BASICS: THE IMPACT OF HIGH-STAKES TESTING ON STUDENT PROFICIENCY IN LOW-STAKES SUBJECTS

Marcus A. Winters, Jay P. Greene
& Julie R. Trivitt

INTRODUCTION

School systems across the nation have adopted policies that reward or sanction particular schools on the basis of their students' performance on standardized tests. Such testing has been a dominant force in education policy since at least the 1990s. More than half the states had already implemented some form of high-stakes test before the No Child Left Behind Act (NCLB) made it universal in 2002. We call a test a high-stakes test when there are meaningful consequences for schools or students that are based on how students perform on the test.

One of the most frequently raised concerns regarding high-stakes testing policies is that they oblige schools to focus on subjects for which they are held accountable but to neglect the rest (Nichols and Berliner 2007; Gunzenhauser 2003; Groves 2002; Patterson 2002; Murillo and Flores 2002; McNeil 2000; Jones et al. 1999). The vast majority of these policies base their rewards or sanctions exclusively on the results of reading and math tests. Though some policies are more expansive than others, few threaten meaningful consequences when students fail to meet standards in subjects such as science, history, or the arts. Failure to assure student mastery of subjects other than basic math and reading could have important implications for the future of human capital in the United States.

If schools reallocate time and resources away from important but low-stakes subjects and toward the high-stakes subjects, with the

result that students achieved in the high-stakes subjects at the expense of proficiency in the low-stakes subjects, we would say that the policy “crowded out” learning in the low-stakes subjects. It is important to note that this definition of crowding out focuses on learning output, not teaching inputs. In other words, if schools increased time spent on math or reading by decreasing time spent on science, we would consider high-stakes testing of math or reading to have crowded out science teaching only if students actually learned less science as a result.

A substantial amount of anecdotal and qualitative evidence suggests that schools and teachers have responded to high-stakes testing by adjusting their teaching styles (McNeil 2000; New York State Education Department 2004) and by shifting focus away from low-stakes subjects (Center on Education Policy 2006; Jones et al. 1999; King and Mathers 1997; Gordon 2002; Groves 2002; Murillo and Flores 2002). However, there is currently very little empirical evidence of the impact of high-stakes testing policies on measured student proficiency in subjects that are not part of the accountability system.

In the only quantitative evaluation of this topic of which we are aware, Jacob (2005) found that Chicago’s high-stakes testing system led to significant learning gains in the low-stakes subjects of science and social studies. However, he found that these gains were smaller than those in the high-stakes subjects of math and reading.

In this paper, we add to the limited previous research by evaluating the effects on student proficiency in the low-stakes subject of science and the high-stakes subjects of math and reading of a high-stakes testing system in Florida that employs sanctions. There are two important reasons to research this question in a system other than Chicago’s. First, by evaluating the impact of sanctions under high-stakes testing on student proficiency in low-stakes subjects in another school system, we can help determine whether the results in Chicago are limited to that area or hold more generally. Second, it is important to investigate outcomes in another system, since some research has found systematic manipulations of Chicago’s high-stakes

exams that could skew results (Jacob 2005; Jacob and Levitt 2003). Previous research in Florida found that the results of that state’s high-stakes exams have not been systematically manipulated and are generally reliable indicators of student proficiency (Greene, Winters, and Forster 2004; West and Peterson 2006).

Florida’s high-stakes testing program is also worth studying because its accountability system, unlike that of many other accountability systems, makes it possible to use a rigorous “regression discontinuity” design, which allows for a causal test of the impact of the program’s sanctions. Beginning in the 2001–02 school year, schools received letter grades reflecting points earned under an elaborate system for capturing several aspects of a school’s performance. As described below, Florida imposes meaningful sanctions only when a school receives a failing grade. We follow the strategy of a previous paper by Rouse et al. (2007) that uses the change in the policy to control for the heterogeneity of schools that receive a failing or passing grade.

We find that students attending schools designated as failing in the prior year made greater gains on the state’s science exam than they would have done if their school had not received the F sanction. The gains that students made in science were similar to those that previous research (which we replicate here) has found that students made in the high-stakes subjects of math and reading. These findings suggest that the incentives of Florida’s high-stakes testing program have not led to significant crowding out of student knowledge in the low-stakes subject of science.

At first, our results may seem counterintuitive, in that high-stakes testing in only certain subjects would be expected to lead schools to focus on those areas. In fact, encouraging schools to shift their priorities toward subjects commonly recognized as academically important (i.e., math and reading) is arguably one of the purposes of the policy.

There are two reasons that high-stakes testing might instead have a positive effect on student achievement in low-stakes subjects. First, the pressure of accountability testing could lead schools to adopt reforms that

improve their overall quality. For example, a school could more effectively motivate its students, or it could improve relations with its teachers. Though schools' purpose may be to improve student scores in math and reading to avoid the sanctions of the high-stakes testing policy, general improvements of this kind might produce across-the-board increases in student achievement. Second, sanctions under high-stakes testing could improve student achievement in low-stakes subjects if the resulting mastery of high-stakes subjects facilitates mastery of other subjects.

Though a true test of the prevalence of either of these kinds of explanations is not available to us, we have discovered evidence suggesting that student proficiency in science has increased under the high-stakes sanctions primarily because the improvements that students have made in math and reading have enhanced their ability to learn science material as well. However, we stress that future research using stronger strategies than are available here to explain a positive relationship between high-stakes testing and student improvement in low-stakes subjects is necessary.

FLORIDA'S A+ ACCOUNTABILITY PROGRAM

Florida is among the nation's leaders in high-stakes testing. Most agree that the state's A+ Accountability Program (A+) is one of the most aggressive programs of its kind. It was clearly a template for the federal NCLB law.

Each year, the state administers a standardized test, the Florida Comprehensive Assessment Test (FCAT), in math and reading to all public school students in the state who are enrolled in grades 3–10. Schools receive letter grades, from A to F, based on the percentage of their students meeting particular achievement levels and the academic progress of students in certain subgroups.

There are two important reasons that we might expect schools deemed to be failing to respond positively. Those that have received an F grade for the first time may be shamed into improving their performance (Fi-

glio and Rouse 2005; Ladd 2001; Carnoy 2001; Harris 2001). Those that have received at least one failing grade may decide to raise their performance because they fear attrition of their student body. This may occur as the result of a policy of issuing Opportunity Scholarships (vouchers) to students in schools that have received two failing grades within a four-year period that they can use to attend another public school or a private school willing to accept the voucher as a full tuition payment.¹ In this paper, we are not particularly concerned with whether these or any other phenomena drive increases in student performance in either high- or low-stakes subjects.

A change in the administration of the program provided an interesting avenue for researching Florida's policy. In the program's initial years, school grades were based on the percentage of students earning level 2 (the second-lowest of five levels) or above on the reading, math, and writing portions of the FCAT and the percentage of eligible students tested. A school could avoid earning an F if at least 50% of tested students scored at achievement level 3 in writing, or if 60% of tested students scored at level 2 in reading or math and 90% of eligible students were tested. If a school met one or two of these criteria, it earned a D. If it met all three of these criteria, it earned a C. Schools with particular subpopulations meeting all three received a B. To earn an A, schools had to meet more stringent requirements for the overall student population and each subpopulation. The opinion was widespread that schools had determined that satisfactory scores in writing were the easiest to achieve under the original school-grading format and that the teaching of writing in struggling schools therefore stressed techniques geared to the writing portion of the exam.

Starting in the 2001–02 school year, Florida adopted an accumulating point system to evaluate schools. Schools earn one point for each percent of students who score in achievement levels 3, 4, or 5 (the three highest of five levels) in reading and one point for each percent of students who score in levels 3, 4, or 5 in math. Schools earn one point for each percent of students scoring 3.5 or above in writing, which is graded from 1 to 6. Schools earn one point for each percent of students who make learning gains in read-

ing and one point for each percent of students who reach a higher achievement level or maintain a 3, 4, or 5 in math. Schools also earn one point for each percent of the lowest-performing readers who make test-score improvements in the year in question. A school that earns fewer than 280 points receives a failing grade. The multifarious nature of the grading process has probably made direct manipulation of the system relatively difficult.

Beginning in the 2002–03 school year, Florida public schools also were required to test for proficiency in science when they administered the math and reading exams. The science part of the FCAT is currently administered to all public school students in grades 5, 8, and 11. The results of the science exam have now been incorporated directly into the accountability program; but during the years of our analysis, they had no effect on the school’s grade, nor did they represent any other form of official accountability.

Several researchers have evaluated the impact of the A+ program on the academic gains of public school students in math and reading (Rouse et al. 2007; Greene and Winters 2004; Chakrabarti 2005; Figlio and Rouse 2005; West and Peterson 2006; Greene 2001). Though there is some disagreement about which aspect of the accountability policy was effective (the threat of vouchers or the shame of an F grade), each of these analyses found that the policy improved the math and reading proficiency of students in public schools designated as failing. We are aware of no previous research analyzing the impact of the A+ program on science test scores.

DATA AND METHOD²

We utilize a data set provided by the Florida Department of Education that contains test scores in math, reading, and science as well as demographic characteristics of the universe of students enrolled in grades 3–10 in Florida public schools. We supplement the individual-level data set with school-level information—specifically, the school’s point total and letter grade under A+ at

the end of the 2001–02 school year. To simplify the comparison of scores in different subjects, we convert the FCAT scores of students who were in our sample into a scale score with a mean of 0 and standard deviation of 1.

In order to align our findings with those in the previous literature, we utilize the comparison strategy implemented in a 2007 study conducted by Rouse et al. that evaluated the impact of Florida’s A+ policy on student achievement in math and reading. Our sample consists of the universe of Florida public school students who were enrolled in the fifth grade in 2002–03 and were promoted at the end of the prior year. This was the first class of fifth-grade students attending a school that had received a letter grade under the revised point system of the A+ policy. We focus on only those students with both a math and reading test score reported in 2001–02 and 2002–03.

We supplement the individual-level data with administrative information on the school’s grade and points earned under the A+ system during the summer of 2002. In the analyses that follow, along with observable characteristics of the student and school we control for both the school’s letter grade at the end of the 2001–02 year and the total points earned under the grading system. The idea here is that controlling for the points earned by the school accounts for differences in school performance, and thus the remaining differences in the performance of students at schools receiving an F grade must reflect responses to the incentives that exist under the accountability policy.

We use this general comparison strategy to perform a series of cross-sectional regressions. We are first concerned with discovering whether students made academic gains in science due to the F sanction, and we also confirm the finding of an impact of the sanction on student proficiency in math and reading. We then evaluate the extent to which any gains made by students in science due to the F sanction were driven by improvements in the overall performance of the school or a symbiotic relationship between learning in the high- and low-stakes subjects.

THE IMPACT OF THE F-GRADE SANCTION ON STUDENT PROFICIENCY IN HIGH- AND LOW-STAKES SUBJECTS

We adopt the strategy of Rouse et al. to measure the impact of the F-grade sanction on student proficiency in science. As a check on our procedure, we also attempt to replicate in math and reading the results of this previous paper.

To evaluate the impact of the F-grade sanction on student proficiency, we estimated cross-sectional regression models using the student's test score on the fifth-grade test in 2002–03 in the subject being evaluated as the dependent variable. The regression controls for a variety of observable characteristics about the student and school, including the letter grade and cubic function³ of the number of points earned by the school at the end of the 2001–02 school year. The variable of interest is that which indicates whether the child's school received an F grade in the prior year.

In the math and reading analyses, we control for a cubic function of the student's test score in that subject at the end of the previous year (when the student was in the fourth grade), which allows us to measure improvements in the student's math and reading proficiency and account for unobserved differences among students. Unfortunately, students did not take a science exam in the fourth grade, so a similar control is not available for the science evaluation. Instead, we use the cubic functions of the students' fourth-grade scores in math and reading to substitute for their scores in science. This procedure assumes that student proficiency in these subjects is highly correlated and that there was no differential relationship in student knowledge among these subjects in the five categories of schools before the F-grade sanction was introduced.

The results of our estimations of student proficiency in math, reading, and science are reported in Table 1. Our findings in math and reading are very similar to those reported by Rouse et al. (2007). Our estimation

Table 1. Regressions evaluating the impact of F-grade sanction on student proficiency and gains in high- and low-stakes subjects

Dependent Var:	Reading		Math		Science	
	Coef.	t	Coef.	t	Coef.	t
Prior Reading	0.759	238.220 ***			0.539	133.430 ***
Prior Reading ^ 2	-0.009	-6.490 ***			-0.008	-5.120 ***
Prior Reading ^ 3	-0.024	-40.290 ***			-0.016	-23.400 ***
Prior Math			0.806	214.900 ***	0.329	79.810 ***
Prior Math ^ 2			-0.004	-2.490 **	0.015	10.480 ***
Prior Math ^ 3			-0.029	-42.980 ***	-0.009	-13.550 ***
A	-0.003	-0.100	0.003	0.060	0.023	0.510
B	-0.005	-0.180	0.005	0.110	0.006	0.160
C	0.006	0.320	0.002	0.090	0.003	0.120
F	0.086	2.940 ***	0.175	3.840 ***	0.087	2.160 **
R-Squared	0.6949		0.6871		0.6588	
N	152,003		152,003		151,604	

Estimated with OLS with robust standard errors clustered by school. Models additionally control for year, limited English-proficiency status, free or reduced-price lunch status, race, gender, disability classification, predicted score in summer of 2002 if the old grading system is kept, a cubic function for the number of points school earned in summer of 2002.

* Statistically significant at 10% level ** Statistically significant at 5% level *** Statistically significant at 1% level

suggests that the scores of students enrolled in an F-graded school exceeded by 0.09 standard deviations in reading and 0.17 standard deviations in math the scores of students in D-graded schools. At the same time, there was no statistically significant difference in the performance of A-, B-, C-, and D-graded schools, strengthening the view that sanctions have an effect on performance. The similarity of our results to those reported by Rouse et al. suggests that we have reproduced their procedure relatively well, which should provide additional confidence in our findings for science.

Column 3 of Table 1 reports the results in science. Here we find that the F-grade sanction produced after one year a gain in student proficiency of about a 0.08 standard deviation relative to students in schools that earned a D grade. The result is significant at the 5% confidence level, meaning that we can be highly confident that the F sanction had a positive impact on students' science test scores.

These findings suggest that the F-grade sanction not only improved student learning in the high-stakes subjects; it also had a positive effect on student proficiency in the low-stakes subject of science. It appears that the positive impact of the F sanction on science proficiency was similar to that found in reading but somewhat lower than that found in math.

UNDERSTANDING THE EFFECTS OF THE F-GRADE SANCTION ON SCIENCE PROFICIENCY

Our finding that the F-grade sanction led to substantial improvements in science proficiency seems odd at first glance. It is clear that under Florida's policy, point-maximizing public schools have an incentive to focus on the high-stakes subjects, even if doing so is to the detriment of the low-stakes subjects. Qualitative research and anecdotal evidence suggest, moreover, that the reallocation of resources precipitated by high-stakes testing has curtailed general student knowledge.

There are two possible ways of explaining how high-stakes testing could increase performance in low-stakes subjects. One is that gains in one subject may facilitate mastery of another. We call this the "correlation effect." Another is that implementation of high-stakes testing could lead to the adoption of policies and attitudes that improve performance generally. For example, high-stakes testing could lead schools to expect improved student achievement across the board, to be shamed into improving their overall performance, to recognize excellence in other subjects, and so on. Rouse et al. find that schools responded to receiving the F-grade sanction in a variety of ways, including lengthening school periods (block scheduling) and increasing time for collaborative planning and class preparation. Such changes in the overall school environment could affect the teaching of science as much as they do the teaching of math or reading. We refer to this possibility as the "systemic effect."

Unfortunately, we cannot produce a true causal estimation of the prevalence of systemic and correlation effects. We can, however, produce some evidence suggesting the relative importance of each by analyzing a regression of student proficiency in science, controlling for observable characteristics used in the previous regression, and including a control for the test-score gain that the student made in math and reading from the 2001–02 to the 2002–03 school year. Here the estimate of the variables for the student's gains in math and reading measures their contribution to gains in science; as explained earlier, this is the correlation effect. The variable for the grade earned by the student's school in 2001–02 measures the grade's impact on science proficiency independently of the correlation effect; this is what we call the "systemic effect."

The results of this estimation are found in Table 2. We find a strongly positive relationship between science scores and gains in math and reading, indicating the likely existence of a correlation effect. The variable representing the independent effect of the F-grade sanction on gains in science is quite small and statistically insignificant, indicating the lack of a systemic effect. These results suggest that the entire gain found in science due to the F-grade sanction is likely due to the correlation effect.

Table 2. Estimating systemic and correlation effect

Dependent Var: Science		
	Coef.	t
Prior Reading	0.644	167.060 ***
Prior Reading ^ 2	0.003	2.190 **
Prior Reading ^ 3	-0.008	-11.840 ***
Prior Math	0.343	81.770 ***
Prior Math ^ 2	0.010	7.050 ***
Prior Math ^ 3	-0.001	-1.770 *
Read Gain	0.424	104.120 ***
Read Gain ^ 2	-0.011	-3.310 ***
Read Gain ^ 3	-0.003	-1.820 *
Math Gain	0.281	60.540 ***
Math Gain ^ 2	0.010	3.270 ***
Math Gain ^ 3	-0.004	-3.080 ***
A	0.023	0.670
B	0.009	0.290
C	0.002	0.080
F	0.001	0.020
R-Squared	0.7371	
N	150,458	

Estimated with OLS with robust standard errors clustered by school. Dependent variable is the student's test score on the fifth-grade science exam. Models additionally control for year, limited English-proficiency status, free or reduced-price lunch status, race, gender, disability classification, predicted score in summer of 2002 if the old grading system is kept, a cubic function for the number of points school earned in summer of 2002.

* Statistically significant at 10% level
 ** Statistically significant at 5% level
 *** Statistically significant at 1% level

SUMMARY AND DISCUSSION

In this paper, we have evaluated whether the F-grade sanction in Florida's A+ program has led schools to increase student learning in the high-stakes subjects of math and reading to the detriment of learning in the important but low-stakes subject of science. Our results indicate that the F-grade sanction led to substantial student gains in the learning of math, reading, and science. Finally, we produced a simple model to explain the impact of high-stakes testing on student learning in low-stakes subjects. We provide some evidence suggesting that virtually all the positive findings in science are attributable to complementarities in the learning of math and reading.

It could be said that student performance in science is not the most authoritative test of the proposition that high-stakes testing crowds out instruction in other subjects, since science proficiency may be more dependent on reading and math proficiency than other subjects are. In effect, such criticism would be assuming the validity of the correlation effect.

Because students in Florida do not take standardized tests in other low-stakes subjects, we are unable to test this hypothesis. It should not be forgotten, however, that much of the discussion of the crowding-out effect focuses on its impact on science learning. Nevertheless, we look forward to future work evaluating the impact of high-stakes testing on student learning in low-stakes subjects other than science.

ENDNOTES

1. The voucher provision of this policy was recently overturned by the state's supreme court, though it was in effect during all the years in which this study takes place.

2. A more technical treatment of the methods utilized in this paper is available online at http://www.manhattan-institute.org/pdf/cr_54_tech_version.pdf.

3. This simply means that we included variables for the points, the points squared, and the points cubed. Use of the cubic function allows for a more flexible model because it relaxes the assumption of linearity in measuring the impact of school points on student proficiency. That is, only controlling for the number of points earned by the school makes the strong assumption that every point has the same impact on science proficiency. That is, it would assume that the impact on a student's science proficiency of a school's raising its score from 100 points to 110 points was identical to the impact of a school's raising its score from 200 to 210 points. The cubic function allows us to account for nonlinearities in this relationship. The same basic argument also holds for the control for a cubic function of the child's prior test scores, which are also discussed and used in this analysis.

- Carnoy, Martin. 2001. "Do School Vouchers Improve Student Performance?" *American Prospect* 12, no. 1 (special report, January): 42–45.
- Center on Education Policy (Washington, D.C.). 2006. "From the Capital [sic] to the Classroom Year 4 of the No Child Left Behind Act." Manuscript. Center. Washington, D.C.
- Chakrabarti, R. 2005. "Do Public Schools Facing Vouchers Behave Strategically?: Evidence from Florida." Manuscript. Program on Education Policy and Governance.
- Figlio, David N., and Cecilia Rouse. 2005. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90 (January): 239–55.
- Gordon, Jenny. 2002. "From Broadway to the ABCs: Making Meaning of Arts Reform in the Age of Accountability." *Educational Foundations* 16, no. 2 (spring): 33–53.
- Greene, Jay P. 2001. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." Manuscript. Manhattan Institute.
- Greene, Jay P., and Marcus A. Winters. 2004. "Competition Passes the Test." *Education Next* 4, no. 3: 66–71.
- Greene, Jay P., Marcus A. Winters, and Greg Forster. 2004. "Testing High-Stakes Tests: Can We Believe the Results of Accountability Tests?" *Teachers College Record* 106, no. 6: 1124–44.
- Groves, Paula. 2002. "'Doesn't It Feel Morbid Here?': High-Stakes Testing and the Widening of the Equity Gap." *Educational Foundations* 16, no. 2 (spring): 15–31.
- Gunzenhauser, Michael G. 2003. "High-Stakes Testing and the Default Philosophy of Education." *Theory into Practice* 42, no. 1 (winter): 51–58.
- Harris, Doug. 2001. "What Caused the Effects of the Florida A+ Program: Ratings or Vouchers?," in *School Vouchers: Examining the Evidence*, ed. Martin Carnoy. Economic Policy Institute.
- Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89, nos. 5–6 (June): 761–96.
- Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118, no. 3 (August): 843–77.
- Jones, M. Gail, Brett D. Jones, Belinda Hardin, and Lisa Chapman. 1999. "The Impact of High-Stakes Testing on Teachers and Students in North Carolina." *Phi Delta Kappan* 81, no. 3 (November): 199.

King, Richard A., and Judith K. Mathers. 1997. "Improving Schools through Performance-Based Accountability and Financial Rewards." *Journal of Education Finance* 23, no. 2 (fall): 147–76.

Ladd, Helen F. 2001. "School-Based Educational Accountability Systems: The Promise and the Pitfalls." *National Tax Journal* 54, no. 2 (June): 385–400.

McNeil, Linda M. 2000. "Creating New Inequalities: Contradictions of Reform." *Phi Delta Kappan* 81, no. 10 (June): 728–34.

Murillo, Enrique G., and Susana Y. Flores. 2002. "Reform by Shame: Managing the Stigma of Labels in High-Stakes Testing." *Educational Foundations* 16, no. 2 (spring): 93–108.

New York State Education Department. 2004. "The Impact of High-Stakes Exams on Students and Teachers." NYSED Policy Brief.

Nichols, Sharon Lynn, and David C. Berliner. 2007. *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Cambridge, Mass.: Harvard Education Press.

Patterson, Jean A. 2002. "Exploring Reform as Symbolism and Expression of Belief." *Educational Foundations* 16, no. 2 (spring): 55–75.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio. 2007. "Feeling the Florida Heat?: How Low-Performing Schools Respond to Voucher and Accountability Pressure." National Center for Analysis of Longitudinal Data in Education Research, Working Paper 13.

West, Martin R., and Paul E. Peterson. 2006. "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments." *Economic Journal* 116, no. 510 (March): C46–62.

NOTES

CENTER FOR CIVIC INNOVATION

Stephen Goldsmith,
Advisory Board Chairman Emeritus
Howard Husock,
Vice President, Policy Research

FELLOWS

Edward Glaeser
Jay P. Greene
George L. Kelling
Edmund J. McMahon
Peter Salins
Fred Siegel
Marcus A. Winters

The Center for Civic Innovation's (CCI) mandate is to improve the quality of life in cities by shaping public policy and enriching public discourse on urban issues. The Center sponsors studies and conferences on issues including education reform, welfare reform, crime reduction, fiscal responsibility, immigration, counter-terrorism policy, housing and development, and prisoner reentry. CCI believes that good government alone cannot guarantee civic health, and that cities thrive only when power and responsibility devolve to the people closest to any problem, whether they are concerned parents, community leaders, or local police.

www.manhattan-institute.org/cci

The Manhattan Institute is a 501(C)(3) nonprofit organization. Contributions are tax-deductible to the fullest extent of the law. EIN #13-2912529