The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects

Jay P. Greene
Endowed Chair; Head
Department of Education Reform
University of Arkansas
jpg@uark.edu

Julie Trivitt
Research Associate
Department of Education Reform
University of Arkansas
jtrivitt@walton.uark.edu

Marcus A. Winters*
Ph.D. Student
Department of Economics
University of Arkansas
winters@uark.edu

* Corresponding Author:
Marcus A. Winters
Manhattan Institute
52 Vanderbilt Ave, 3rd Floor
New York, NY 10017
Phone: 646-839-3354

JEL Classification: H11; I28; J24

Keywords: educational economics, human capital, productivity, high-stakes testing, accountability

Acknowledgements: We would like to thank Bruce Dixon, Robert Costrell and Ryan Marsh for their valuable comments. All remaining errors are our own.

Abstract:

An important criticism of high-stakes testing policies – policies that reward or sanction schools based on their students' performance on standardized tests – is that they provide schools with an incentive to focus on those subjects that play a role in the accountability system while decreasing attention to those subjects that are not part of the program. This paper evaluates the impact of Florida's high-stakes testing policy on student proficiency in the low-stakes subject of science. We confirm that students in schools that were sanctioned under the policy made substantial gains in the high-stakes subjects of math and reading. We also find that students in these schools made similar achievement gains in the low-stakes subject of science. Further, we find that the relationship between student learning in high-stakes subjects and science was the same for students in sanctioned and unsanctioned schools, indicating that the policy did not "crowd out" learning in the low-stakes subject. We then present a simple model to illustrate how high-stakes testing could have an overall positive impact on student knowledge in low-stakes subjects. We provide a limited test for this relationship and find some suggestive evidence that the increase in student learning in the low-stakes subject could entirely be due to correlations in the learning across the high- and low-stakes subjects.

Introduction

School systems across the nation have adopted accountability policies that reward or sanction schools based on their students' performance on standardized tests. Such high-stakes 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.

One of the most frequently raised concerns regarding high-stakes testing policies is that they provide schools with an incentive to focus on those subjects that play a role in the accountability system while decreasing attention to those subjects for which student proficiency is not part of the program (Nichols and Berliner 2007; Gunzenhauser 2003; Groves 2002; Patterson 2002; Murillo 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 hold meaningful consequences for whether students meet standards in other subjects such as science, history, or the arts. Failing to provide students with skills in subjects other than basic math and reading could have important consequences for the future of human capital in the United States.

Schools have a clear incentive to reallocate time and resources away from these arguably important but low-stakes subjects and toward the high-stakes subjects. If such a reallocation led to increased student achievement 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 subject. It is important to emphasize that this definition of crowding out focuses on learning output, not teaching inputs. That is, if schools responded to high stakes testing by increasing time spent on math or reading while decreasing time spent on science we would only consider this

crowding out meaningful if students learned relatively less science material due to the decreased attention to the subject. Importantly for our purposes, any such crowding out effect clearly implies a change in the relationship between student knowledge in the high- and low-stakes subjects.

A substantial amount of anecdotal and qualitative evidence indicates 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 removing focus from low-stakes subjects (CEP 2006; Jones et.al.1999; Gunzenhauser & Noblit 2001; King & Mathers 1997; Gordon 2002; Groves 2002; Murillo & Flores 2002). However, there is currently very little empirical evidence on 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 (2004) finds that Chicago's high-stakes testing system led to significant additional learning gains in the low-stakes subjects of science and social studies. However, he finds that these gains in low-stakes subjects due to the policy 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 impact of the receipt of sanctions under high-stakes testing in Florida on the relationship between student proficiency in the low-stakes subject of science and the high-stakes subjects of math and reading. There are two important reasons to research this question in other school systems. By evaluating the impact of sanctions under high-stakes testing on student proficiency in low-stakes subjects in another school system we can determine whether the results from Chicago are specific to that area or hold more generally. Further, the systematic manipulations of Chicago's high-stakes

exam found in previous research (Jacob 2004; Jacob and Levitt 2003) may be driving the differential relationship between student learning gains in the high-stakes and low-stakes in a way that may not exist in Florida. Previous research in Florida finds that the results from that state's high-stakes exam 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 policy is also interesting because its design allows for use of a regression discontinuity type approach not available in many other accountability systems. As described below, whether public schools in Florida receive a meaningful sanction from the accountability policy depends on whether they receive an F grade from the state. Beginning in the 2001-02 school year, schools received letter grades by earning points under an elaborate system designed to incorporate several aspects of the school's performance. 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.

Our findings suggest that the differential incentives under Florida's policy have had no significant impact on the relationship between student learning in the high-stakes subjects of math and reading and the low-stakes subject of science. We then go on to confirm the findings of Rouse et al. that receiving an F-grade led to increased student proficiency in math and reading. Furthermore, we find that students in this cohort made similar gains on the state's science exam as they did on the math and reading exams. Thus, our results indicate that Florida's policy has not led to significant crowding out of the learning of science and that this finding is not due to failure of the policy to produce student gains in academic proficiency.

At first, these results may seem counterintuitive. In some ways, it seems obvious that high-stakes testing in certain subjects would lead schools to focus on those areas. In fact,

encouraging schools to shift their priorities toward those topics more conventionally recognized as academically important (i.e. math and reading) is arguably one of the explicit purposes of the policy. To reconcile our findings with incentive maximizing behavior we present a simple model that illustrates how high-stakes testing could have an overall positive impact on student knowledge in low-stakes subjects. The framework allows for potential complementarities in learning across subjects and the possibility of a positive effect from testing on the overall school environment. We then take this theoretical model to the data and find some limited evidence that the increase in student learning in science is primarily due to correlations with the learning of math, and reading.

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. This policy was a clear template for the structure of 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 potentially important treatments for schools that receive an F grade under the program. First, students in schools that receive two failing grades from the state within the previous four-year period are offered Opportunity Scholarships (vouchers) that they can use to attend another public school or a private school that is wiling to accept the voucher as a full

tuition payment.[1] After receiving their first F grade, schools may attempt to increase their performance in order to avoid receiving a second F and thus losing students and funding to the voucher program. Secondly, the act of distinguishing schools as "failing" could have a motivational effect on schools. In particular, several researchers have suggested that such policies are effective because they "shame" schools into better performance (Figlio & Rouse 2005; Ladd 2001, Carnoy 2001, Harris 2001). In this paper we are not particularly concerned with which of these or any other aspects drive increases in student performance in either the high or low-stakes subjects.

A change in the administration of the program provides an interesting avenue for researching Florida's policy. In the initial years of the program, 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. Schools 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 are tested. Schools that met all three of these criteria earned a C and schools meeting them for all specific sub-populations received a B. To earn an A schools had to meet more stringent requirements for the overall student population and each specific sub-population. The widespread opinion was that schools determined the writing requirement was the easiest to achieve under the original school grading format and struggling schools began teaching specific writing techniques geared to that portion of the exam to avoid earning an F.

In 2001-02 the grading criteria were changed to evaluate schools based on a variety of characteristics using an accumulating point system. Schools earn one point for each percent of

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

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 reading and one point for each percent of students who make learning gains by improving achievement levels, or maintain a relatively high level of 3, 4, or 5 in math. Schools also earn one point for each percent of the lowest performing readers who make learning gains from the previous year. A school that earns fewer than 280 points receives a failing grade. The complicated nature of the grading process should make direct manipulation of the system relatively difficult.

Beginning in the 2002-03 school year, Florida public schools also were required to administer a science version of the FCAT when they administered the math and reading exams. The science test is currently administered to all public school students in grades 5, 8, and 11. Though the results of the science test are made publicly available, they have no effect on the school's grade, nor do they have any other formal accountability purpose.

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 2005; Greene 2001). Though there is some disagreement about which aspect of the accountability policy was effective (threat of vouchers or shaming), each of these analyses found that the policy has improved the performance 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

We utilize a dataset provided by the Florida Department of Education that contains test scores in math, reading, and science as well as demographic characteristics for the universe of students enrolled in grades 3 – 10 in a Florida public schools. We supplement the individual-level dataset with school-level information about the school's point total and letter grade under A+ at the end of school year 2001-02. To aid in score comparisons across subjects, we convert the FCAT scores into a scale score with a mean of 0 and standard deviation of 1 for students who were in our sample.

For this analysis we focus on a sub-set of students to allow us to estimate the impact of the A+ grading program. Following Rouse et al. (2007), our analysis includes only those students who were in the fifth grade in 2002-03 and were cumulatively promoted at the end of the 2001-02 school year. The class of 2002-03 is the first to attend a school after it had been assigned a grade under the revised point system. Also, for our particular purposes we are interested only in fifth grade students because this is the only elementary grade in which a science exam was administered.

Method

In order to best align our findings with the previous literature, we utilize the basic comparison strategy implemented in a recent previous study evaluating the impact of Florida's A+ policy on student achievement in math and reading. In particular, we attempt to reproduce the basic procedure adopted by Rouse et al. (2007).

Our sample consists of the universe of Florida public school students who were enrolled in the fifth grade in 2002-03 and were cumulatively promoted at the end of the prior year. This was the first class of fifth grade students attending a school after it had received a letter grade

under the revised point system of the A+ policy. We further focus 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 we control both for 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 after controlling for the points earned by the school accounts for differences in school performance, and thus any additional impact of receiving an F-grade is the causal impact of the pressure faced under the accountability policy.

Following Rouse et al. we also calculate and control for the grade that the public school would have received during the summer of 2002 if the state had not changed its grading policy that year. Since the change in the grading policy was relatively sudden, accounting for the grade that the school would have received under the old system could help to account for the extent to which schools could have been surprised by the policy.[2]

We use this general comparison strategy to perform a series of cross-sectional regressions. We are first concerned with discovering whether there is a differential relationship between student knowledge in the high-stakes subjects and science. We then evaluate whether any such differential relationship (or lack there of) is caused by differential gains in student learning in these subjects between the fourth and fifth grade years.

The Impact of High-stakes Testing on Relative Knowledge in High and Low-stakes Subjects

[2] Our calculation of predicted 2002 school grades under the old system are similar to those of Rouse et al. for grade levels A, B, and F. However, several schools that our results suggest would have been classified as receiving a D grade were classified as they would have received a C grade by Rouse et al. The reason for this difference is unclear. However, as we will see, this seems not to produce very different results in the results of our estimations.

The idea that high-stakes testing causes schools to crowd out learning in low-stakes subjects implies that the relationship between student knowledge in high- and low-stakes subjects changes when faced with high-stakes sanctions. We assume that knowledge across subjects is significantly but not perfectly correlated. If schools faced with high-stakes sanctions respond by increasing student proficiency in math and reading to the detriment of their learning in science then we should see a difference in the relationship between student knowledge in the high- and low-stakes subject in schools subject to the sanction. That is, in a sanctioned school we would expect that each unit of knowledge in math or reading would translate into fewer units of knowledge in science.

We can directly estimate whether obtaining an F-grade causes a differential relationship between student knowledge in the low-stakes subject of science and the high-stakes subjects of math and reading by utilizing student test score and demographic information from the 2002-03 school year. Our basic model for estimation takes the form:

$$Science_{ist} = \beta_0 + \beta_1 Grade_{st-1} + \beta_2 X_{ist} + \beta_3 f(POINTS_{st-1}) + \beta_4 Y_{ist} + \beta_5 (Grade_{st-1} * Y_{ist}) + \varepsilon_{ist} \quad (1)$$

Where $Science_{ist}$ is the test score in science for student i in school s during year t; Grade is the letter grade earned by the student's school in the summer of 2002; X is a series of control variables for student attributes; *f*(POINTS) is a cubic of the number of points earned by the school during the summer of 2002; Y is the student's score on the high-stakes math or reading exam; and ε is a stochastic error term clustered at the school level.

We can estimate equation (1) efficiently using OLS. For our current consideration, the primary variable of interest is β_5, which can be interpreted as the differential relationship between student proficiency in science and proficiency in math and reading in schools that

earned an F the previous year compared to students in those schools that operated without the F grade sanction. If the sanction of earning an F grade leads schools to crowd out learning in science then we would expect these coefficients to be significantly negative, indicating that each additional unit of knowledge in math or reading represents fewer units of knowledge in science in F graded schools.

The results of estimating multiple versions of equation (1) are reported in Table 1. In the table we report results of estimation incorporating math and reading scores individually and then together in the equation.

[TABLE 1 ABOUT HERE]

The first result to notice is that our assumption that student knowledge in math, reading, and science is correlated appears to be accurate. Each of our estimations finds a statistically significant relationship between student test scores in science and scores in math and reading.

The combined results of the all the estimations indicate that earning an F grade has no significant impact on the relative knowledge of students in science and proficiency in reading. In none of the reported estimations is the coefficient on the F grade or interaction between reading scores and earning any particular grade found to be statistically significant at any conventional level.

The results in math are somewhat more interesting. Here we pay closest attention to the regression including both student proficiency in math and reading. The comparison group for the school grade and interaction terms are students in a school that received a D-grade. The insignificant t-statistics on the F-grade and C-grade interaction variables indicates that there is no statistical difference in the relationship between student proficiency in math and science for

students in F, C, and and D-graded schools.[3] However, the results do suggest that students in schools graded A, or B had a differential relationship between math and science proficiency than students in schools graded C, D, or F. Further, an F-test of the coefficients finds a significant differential relationship between math and science proficiency among the A, and B schools themselves. It is worth noting, however, that these differences are relatively minor.

These results suggest that there has been no crowding out of science for math proficiency in the F, C, and D schools relative to each other, but that there has been a small but statistically significant crowding-out of math in these lower performing schools relative to the higher performing A, and B schools. This is particularly interesting because we would most expect to find a difference in the response of F relative to D schools because the distinction between these grades is the distinction between receiving or not receiving a sanction. One possibility is that D and C schools respond to their close-call from receiving the F-grade sanction in the same way that F-schools respond to the true sanction. Future research is necessary to shed greater light on this issue.

The Impact of the F Grade Sanction on Student Gains in High and Low-stakes Subjects

These results suggest that sanctions under Florida's high-stakes testing program have not led to substantial crowding out of student proficiency in science. One possible reason for the lack of such a change in this relationship would be if the F grade sanction did not lead to increased student proficiency in any subject. That is, perhaps the high stakes testing program had no effect at all.

Several previous studies have found that the F grade sanction in Florida has had a positive impact on student proficiency in math and reading. Here we attempt to reproduce the

[3] An F-test confirms that there is no statistical difference between the interaction with the C-grade and F-graded schools at the 10% level.

findings of Rouse et al. that the F-grade sanction led to substantial improvements in student math and reading proficiency. Further, we go on to provide the first evaluation of the impact of the F-grade sanction on student learning gains in science.

To evaluate student gains in these subjects we follow Rouse et. al. to estimate a simple education production function taking the form:

$$T_{ist} = \delta_0 + \delta_1 f(T_{ist-1}) + \delta_2 Grade_{st} + \delta_3 X_{ist} + \delta_4 f(POINTS_{st-1}) + \rho_{ist} \quad (2)$$

Where T indicates the student's test score; $f(T_{ist-1})$ is a cubic of the student's test score in the prior year (fourth grade); ρ is a stochastic error term clustered by school; and all other variables are as previously defined.

In the case of math and reading, we can directly estimate (2) by OLS. However, the situation is more difficult in the case of science because schools only test science in the fifth grade. Thus, we are unable to directly observe $Y_{i,a,s,t-1}$ for science and must develop a proxy for its value. Instead, we use the student's prior achievement in math and reading to proxy for prior proficiency in science. This assumes that student proficiency in science is highly correlated with proficiency in math and reading, which is consistent with our findings reported in Table 1. We further assume that there was no systematic difference in the relationship between student proficiency in these subjects in schools that did and did not receive an F-grade in the summer of 2002 prior to the school receiving that grade.

The results of our estimations of (2) in math, reading, and science are reported in Table 2. Our findings in math and reading are very similar (well within a 95 percent confidence interval of their original findings) of those reported by Rouse et al. (2007). Our estimation suggests that students enrolled in an F-graded school made gains of 0.09 standard deviations in reading and 0.17 standard deviations in math relative to students in D-graded schools. Further, there is no

statistically significant difference in the performance of A, B, C, and D-graded schools. The similarity of our results with those reported by Rouse et al. suggests that we have reproduced their procedure relatively well, which provides additional confidence to our results in science.

[TABLE 2 ABOUT HERE]

Column 3 of the table reports the results in science. Here we find that the F grade sanction also had a substantial positive impact on the gains made in student proficiency in science of about a 0.08 standard deviation gain after one year. This impact is quite similar to that in reading but lower than that in math. The result is significant at the 5% confidence level.

These results are consistent with the previous findings on the relationship between student learning in the high-stakes subjects and learning in science. Rather than being driven by a lack of gains in math and reading, the prior result appears to be driven by students making substantial but *similar* gains in the high-stakes subjects and science. We appear to have found no difference in the relationship between student learning in reading and science because students made essentially identical gains in these subjects. Further, our limited finding of a potential differential relationship between math and science proficiency is driven by students making particularly large gains in math. That is, students in F-graded schools made greater gains in science than their contemporaries in unsanctioned schools but these gains were slightly lower than the gains made in the high-stakes subject of math.

Understanding the Effects of the F Grade Sanction on Science Proficiency

The combination of our findings that the F-grade sanction led to substantial improvements in science proficiency and that science was not substantially crowded out by additional learning in math or reading may seem quite odd. It is clear that under Florida's policy, point maximizing public schools have an incentive to focus on the high-stakes subjects to the

detriment of the low-stakes subjects in the classroom. This somewhat obvious relationship is the driving force behind the finding of such crowding out in the large amount of qualitative research and anecdotal evidence on the impact of high-stakes testing on general student knowledge.

There are, however, two potential reasons that high-stakes testing could increase performance in low-stakes subjects which are often ignored in such discussions. First, there may be correlations in knowledge across different subjects such that if students gain additional knowledge in one subject (math or reading) this transforms into the ability to learn in other subjects as well (science). Second, implementation of high-stakes testing could lead schools to adopt policies and attitudes that improve their performance school-wide, not only in the high-stakes subjects. For example, high-stakes testing could lead schools to expect high achievement for students generally, shame schools into improving their overall performance, lead principals to adopt policies that recognize excellence across the school, etc. Rouse et al. (2007) find that schools responded to receiving the F-grade sanction in a variety of ways, including adopting block scheduling and increasing time for collaborative planning and class preparation. Such changes in the overall school environment could have similar effects in the teaching of science as in math or reading.

We can incorporate what we have learned from our previous estimations to develop a theoretical model for the impact of adopting high-stakes testing on changes in student proficiency in low-stakes subjects. Consider a linear model for learning in science similar but, for illustration purposes, more specific to that of equation (1):

$$\begin{aligned} Science_t &= \psi_0 + \psi_1 Science_{t-1} + \psi_2 X_{ist} + \\ &\psi_3[Math_t(F) - Math_{t-1}] + \psi_4[\mathrm{Re}\,ad_t(F) - \mathrm{Re}\,ad_{t-1}] + \psi_5 F_{ist-1} + \mu_{ist} \end{aligned} \quad (4)$$

Where Science, Math, and Read indicate the student's proficiency in those respective subjects, F indicates that the student attended a school that received an F-grade, and X continues to be student demographic controls. This revision of equation (1) directly incorporates our finding that science, math, and reading proficiency are all partially functions of the F grade sanction. Further, we only allow student proficiency in year t to be a function of the F-grade to evaluate the impact of changing the policy, which is also the focus of our empirical approach.

This very simple illustrative model also allows student learning in science to be driven by learning in other fundamental subjects. This relationship suggests improvements in student knowledge in alternative subjects are complimented by learning in the core subjects of math and reading. That is, a student who learns more in math and/or reading could gain additional basic skills that make them better able to acquire knowledge in another subject, such as science.

We can evaluate the impact of earning an F grade on science proficiency by taking the partial derivative of (4) with respect to F:

$$\frac{\delta Science_t}{\delta F} = \underbrace{\psi_4 \frac{\delta Math_t}{\delta F} + \psi_5 \frac{\delta \mathrm{Re}ad_t}{\delta F}}_{CorrelationEffect} + \underbrace{\psi_6}_{SystemicEffect} \qquad (5)$$

Equation (5) separates the impact on science proficiency occurring from the sanction from earning an F-grade into two factors. The first set of factors is the indirect effect of earning an F-grade. This comes from the increased proficiency in high-stakes subjects multiplied by the degree of complementarities between gaining a unit of knowledge in these subjects on the student's knowledge in the low-stakes subject. We refer to this impact as the "correlation effect" because it derives from the correlation in learning across subjects. This effect incorporates the fact that there are complements in student learning across subjects.

The second term is the direct effect of earning an F grade on the student proficiency in the low-stakes subject. The impact of this term could be either positive or negative. If the F-grade sanction tends to decrease student learning in science due to crowding out or any other factors then the coefficient will be driven downward. However, it is also conceivable that the F-grade sanction could have indirect positive consequences for science proficiency if schools adopt structural reforms that increase the productivity of the entire school. The expected sign of Ψ_6 is thus ambiguous. We refer to this as the "systemic effect" because it results from the response of the overall school system to earning an F-grade.

The results of the previous evaluations in this paper allow us to provide a suggestive evaluation of whether the extent to which the positive impact of earning an F grade on science proficiency is driven by a correlation or systemic effect. However, it is worth emphasizing that the procedures pursued below should not be considered true causal tests of the existence of a correlation or systemic effect on science achievement. Rather, we only seek to pursue our data to the best of its abilities to get a suggestive account for what might be driving the results. Future research utilizing alternative strategies is likely necessary for firm findings on this question.

The coefficients ψ_4 and ψ_5 are the impact of a unit of student proficiency in math and reading, respectively, on proficiency in science. We have consistent estimates of these values from estimation of (1) without inclusion of the F and interaction terms (since they are later found to be equal to zero) found in Column (1) of Table 1. Further, the coefficient on φ_5 in our estimations of (2) reported in Columns (1B), (2B), and (3B) in Table 2 are consistent for $\frac{\delta Science}{\delta F}$, $\frac{\delta Math}{\delta F}$, and $\frac{\delta \mathrm{Re}ad}{\delta F}$, respectively.

We can substitute our consistent estimates for these values into (5). This leaves only ψ_6 to be determined, which we can do directly since we are left with a single equation with one

unknown variable. After this substitution and algebra we find that $\psi_6 = -0.023$. This indicates that the large portion of the increase in science scores due to the F-grade is caused by the correlation effect, but that there is some potential for a small crowding out systemic effect. However, this algebraic procedure allows for no significance testing of the estimate.

To evaluate the robustness of this result and provide significance testing we go on to directly estimate (4) from the data, incorporating a stochastic error term clustered by school. We continue to proxy for $Science_{t-1}$ using prior student achievement in math and reading. Results of this estimation are found in Table 3. The estimation finds a strongly positive relationship between science scores and gains in math and reading, indicating the likely existence of a correlation effect. We also see that the coefficient evaluating the independent effect of the F-grade sanction on gains in science is quite small, and here is statistically insignificant at any conventional level. These results again indicate that the entire gain found in science due to the F-grade sanction is likely due to the correlation effect.

[TABLE 3 ABOUT HERE]

This finding is also consistent with the results from the estimations of (1) found in Table 1. In those estimations we found that earning an F grade had no independent effect on the student's level of science proficiency. The above procedure simply incorporates those results into the discussion of test score gains as well.

Summary and Discussion

In this paper we have evaluated whether the F grade sanction of Florida's A+ program has led schools to increase student learning in 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

and that science gains were not substantially decreased from crowding out due to the incentive to focus on high-stakes subjects. Finally, we produced a simple model to potentially explain the impact of high-stakes testing on student learning in low-stakes subjects. We provide some suggestive evidence that virtually all of the positive findings in science could potentially be attributed to complementarities in the learning of math, reading, and science.

One reasonable criticism of our findings here is that we may expect that student proficiency in science is more highly correlated with learning in math and reading than other low-stakes subjects. If this is the case then the fact that the totality of the gains in science occurring from the F grade sanction are driven by the correlation effect would imply that high-stakes testing would in fact crowd out learning on such other subjects where knowledge is less correlated with math or reading.

Unfortunately, students in Florida do not take standardized tests in other low-stakes subjects so we are unable to test this hypothesis. Future research in other areas with high-stakes testing that do also test in multiple low-stakes subjects would be useful to test such an effect. However, much of the discussion about the crowding out effect of high-stakes testing focuses on its impact on science learning. Thus, though this study is not a complete evaluation of the impact of high-stakes testing in knowledge in *all* low-stakes subjects, we do evaluate a subject that is thought to be important by practitioners, researchers and policymakers. We look forward to future work evaluating the impact of high-stakes testing on student learning in low-stakes subjects other than science.

References

Angrist, Joshua D., and Victor Lavy. 1999. “Using maimonides' rule to estimate the effect of class size on scholastic achievement.” *Quarterly Journal of Economics* 114, (2) (May): 533-575.

Carnoy, Martin. 2001. “Do school vouchers improve student performance?” *American Prospect* 12, (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, The Center [database online]. Washington, D.C.

Chakrabarti, R. 2005. “Do public schools facing voucher behave strategically? Evidence from Florida.” Manuscript. Program on Education Policy and Governance.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. “The central role of noise in evaluating interventions that use test scores to rank schools.” *American Economic Review* 95, (4) (September): 1237.

Figlio, David N., and Cecilia Rouse. 2005. “Do Accountability and Voucher Threats Improve Low-Performing Schools?” *Journal of Public Economics* 90, (January) 239-255.

Gordon, Jenny. 2002. “From broadway to the ABCs: Making meaning of arts reform in the age of accountability.” *Educational Foundations* 16, (2) (Spr 2002): 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. 2007. “Revisiting Grade Retention: An evaluation of Florida’s test-based promotion policy.” *Education Finance and Policy*, Volume 2, Number 4.

Greene, Jay P. and Marcus A. Winters. 2004. “Competition Passes the Test.” *Education Next* 4, (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* Volume 106 Number 6, 2004: 1124-1144

Groves, Paula. 2002. “’Doesn't it feel morbid here?’ high-stakes testing and the widening of the equity gap.” *Educational Foundations* 16, (2) (Spr 2002): 15-31.

Gunzenhauser, Michael G. 2003. “High-stakes testing and the default philosophy of education.” *Theory into Practice* 42, (1) (Win 2003): 51-58.

Harris, Doug. 2001. “What Caused the Effects of the Florida A+ Program: Ratings or Vouchers?” in *School Vouchers: Examining the Evidence*, edited by 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, (5-6) (June): 761-796.

———. 2004. “Public housing, housing vouchers, and student achievement: Evidence from public housing demolitions in Chicago.” *American Economic Review* 94, (1) (March): 233-258.

Jacob, Brian A., and Lars Lefgren. 2004. “The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago.” *Journal of Human Resources* 39, (1) (Winter): 50-79.

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, (3) (August): 843-877.

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, (3) (Nov): 199.

King, Richard A., and Judith K. Mathers. 1997. “Improving schools through performance-based accountability and financial rewards.” *Journal of Education Finance* 23, (2) (Fall 1997): 147-76.

Ladd, Helen F. 2001. “School-based educational accountability systems: The promise and the pitfalls.” *National Tax Journal* 54, (2) (June): 385-400.

McNeil, Linda M. 2000. “Creating new inequalities: Contradictions of reform.” *Phi Delta Kappan* 81, (10) (Jun 2000): 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, (2): 93-108.

New York State Education Department 2004. “The impact of high-stakes exams on students and teachers.” NYSED Policy Brief.

Nichols, Sharon Lynn. Berliner,David C. 2007. *Collateral damage : How high-stakes testing corrupts America’s schools*. Cambridge, MA: Harvard Education Press.

Patterson, Jean A. 2002. “Exploring reform as symbolism and expression of belief.” *Educational Foundations* 16, (2) (Spr 2002): 55-75.

Rouse, C.E., Hannaway, J., Goldhaber, D., and Figlio, D. (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.

van der Klaauw, Wilbert. 2002. “Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach.” *International Economic Review* 43, (4) (November): 1249-1287.

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, (510) (March): C46-62.

Table 1
Regressions evaluating the impact of F-grade sanction on relative student knowledge in high and low-stakes subjects

	Coef.	t		Coef.	t		Coef.	t	
Reading	0.761	70.350	***				0.555	44.200	***
Math				0.629	66.920	***	0.329	29.170	***
A	0.013	0.390		0.008	0.180		0.011	0.340	
B	0.012	0.430		0.003	0.080		0.007	0.240	
C	0.005	0.280		0.007	0.280		0.006	0.320	
F	0.020	0.570		-0.017	-0.450		-0.013	-0.390	
A * Read	0.014	1.180					-0.007	-0.520	
B * Read	0.015	1.210					-0.001	-0.050	
C * Read	0.012	1.000					0.014	1.020	
F * Read	0.002	0.090					0.025	0.830	
A * Math				0.071	6.760	***	0.046	3.750	***
B * Math				0.048	4.240	***	0.030	2.300	**
C * Math				0.022	2.020	**	0.002	0.180	
F * Math				-0.029	-1.170		-0.035	-1.350	
R-Squared	0.6759			0.6143			0.7231		
N	150,458			150,458			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 Priced Lunch status, race, gender, disability classification, predicted score in summer of 2002 if kept the old grading system, a cubic 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

Table 2
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 Priced Lunch status, race, gender, disability classification, predicted score in summer of 2002 if kept the old grading system, a cubic 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

Table 3
Estimating Systemic and Correlation Effect

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 Priced Lunch status, race, gender, disability classification, predicted score in summer of 2002 if kept the old grading system, a cubic 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