NBER WORKING PAPER SERIES

TESTING, CRIME AND PUNISHMENT

David N. Figlio

Working Paper 11194
http://www.nber.org/papers/w11194

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2005

I am grateful to the set of undisclosed school districts for providing me with the data to conduct this project, and to the National Science Foundation, the National Institutes of Child Health and Human Development, and several private foundations for research support. I benefited from conversations with Eric Brunner, Julie Cullen, Randy Eberts, Hendrik Juerges, Larry Kenny, Jens Ludwig, Steve Raphael, Lorien Rice, and participants at the AEA, APPAM and IIPF meetings and seminars at Chicago, Florida, Georgia State and the National Bureau of Economic Research, as well as the comments of an anonymous reviewer. I alone am responsible for errors.The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Testing, Crime and Punishment
David N. Figlio
NBER Working Paper No. 11193
March 2005
JEL No. I2

ABSTRACT

The recent passage of the No Child Left Behind Act of 2001 solidified a national trend toward increased student testing for the purpose of evaluating public schools. This new environment for schools provides strong incentives for schools to alter the ways in which they deliver educational services. This paper investigates whether schools may employ discipline for misbehavior as a tool to bolster aggregate test performance. To do so, this paper utilizes an extraordinary dataset constructed from the school district administrative records of a subset of the school districts in Florida during the four years surrounding the introduction of a high-stakes testing regime. It compare the suspensions of students involved in each of the 41,803 incidents in which two students were suspended and where prior year test scores for both students are observed. While schools always tend to assign harsher punishments to low-performing students than to high-performing students throughout the year, this gap grows substantially during the testing window. Moreover, this testing window-related gap is only observed for students in testing grades. In summary, schools apparent act on the incentive to re-shape the testing pool through selective discipline in response to accountability pressures.

David N. Figlio
Department of Economics
University of Florida
Gainesville, FL 32611-7140
and NBER
figlio@ufl.edu

Testing, Crime and Punishment

Introduction

The recent passage of the No Child Left Behind Act of 2001 solidified a national trend toward increased student testing for the purpose of evaluating public schools. Under the new federal law, states must develop and administer rigorous curriculum-based assessments to every student in grades three through eight in every year. These tests must be used to evaluate schools, and in the case of the many schools receiving federal Title I aid, aggregate student performance on these examinations will be associated with substantial rewards and sanctions, including redirection of funding to provide for school choice and privately-provided supplemental services, and ultimately potential replacement of school leadership and staff or state takeover of operations.

This new environment for schools provides strong incentives for schools to alter the ways in which they deliver educational services. Indeed, this is the rationale behind the school accountability movement. Schools may, for instance, respond to these incentives by focusing additional attention on the curricular content of the examinations, or may experiment with innovative methods of instructional delivery. On the other hand, schools might also respond to incentives to improve by making choices that serve to reduce the informative value of the aggregate test score signal as an indicator of aggregate student achievement. One possibility is so-called "teaching to the test," in which schools focus on test-preparation skills and tailor their instruction to subjects included on the examination with high probability. While controversial, it is unclear as to whether teaching to the test is desirable or undesirable, especially when the test content is rigorous and wide-ranging. Another example of behavior that could tend to reduce the

informative signal of aggregate test scores involves the assignment of students to special education. Several recent authors, including Cullen and Reback (2002), Figlio and Getzler (2002) and Jacob (forthcoming), have shown that schools tend to respond to accountability systems and testing regimes by classifying more marginal students as disabled. Jacob (forthcoming) also finds evidence of systematic grade retention in response to test-based accountability. One interpretation of these results is that schools are behaving in an insidious manner, reclassifying potentially low-performing students into test-excluded categories in order to make average test scores look better. But it is also unclear whether this behavior is desirable or undesirable, given that one could legitimately make the argument that rather than "gaming the system," this pattern reflects an increased attention to assessment associated with the testing regime, and students who previously may have slipped through the cracks are now being appropriately classified. Figlio and Winicki (forthcoming) demonstrate that schools faced with accountability pressures respond by systematically changing school feeding programs in order to provide students with short-term nutritional advantages on test days. Here too, one can tell stories for whether this is a desirable or undesirable outcome of high-stakes testing.

This paper explores an entirely different type of response to the introduction of testing regimes. I investigate whether schools may employ discipline for misbehavior as a tool to bolster aggregate test performance. The mechanism through which discipline can assist in this endeavor is straightforward: During the testing window, potentially low-performing students could be given harsher punishments (longer suspensions) than potentially high-performing students receive for similar infractions, because the school may desire to have as many high-performing students as possible in school to take the

examination but at the same time hopes to have more low-performing students stay home during testing periods. Students receiving long suspensions during the testing window are much more likely to miss the examination and its make-up dates: Students suspended for a week or more during the testing window are twice as likely to miss their reading and mathematics examinations as are students who are not suspended for a week or more during this window, and students suspended for two or more weeks during the testing window are 2.6 times as likely to miss the reading examination and 2.5 times as likely to miss the mathematics examination as are students not suspended for this duration. Therefore, suspended students comprise a very large share of the students who do not take the test. Schools across the performance level distribution have the incentive to use discipline as a tool if there exists a gradation of school report card "grades." For instance, in Florida schools are currently explicitly graded on a scale from A to F, and were previously evaluated on a four-point scale. With multiple possible grading categories, more schools are on the margin of a different performance grade level.

To investigate whether schools employ discipline in an apparent attempt to "game the testing system," I utilize an extraordinary dataset constructed from the school district administrative records of a subset of the school districts in Florida. (For confidentiality reasons, these school districts must remain unidentified in this research. As a result, I also cannot reveal identifying details on the demographic attributes of the school districts in question.) This dataset provide information on every disciplinary suspension, both in-school and out-of-school, during the four school years from 1996-97 through 1999-2000, the first four years following the introduction of the Florida Comprehensive Assessment Test (FCAT), Florida's high-stakes examination used to evaluate schools. In the first

school year, the test was administered, but the test scores were viewed as preliminary, with scores neither publicly-reported nor used for school accountability. In these data, I can identify the students involved in each incident, and can therefore match them with demographic and test score records. Most importantly, I know the specific timing of the suspensions imposed, so I can compare suspension durations over the testing cycle. I compare the disciplinary actions taken against two students suspended for the same incident, and explore whether, after controlling for incident fixed effects, the suspensions meted out to each student are related to their prior year's test scores in the manner described above. In addition, since Florida only had high-stakes testing of students in a subset of grades (four, five, eight and ten) during the study period, I can also ascertain whether the discipline-prior test score relationship over the testing cycle is different in high-stakes testing grades than in other grades.

In all, I compare the suspensions of students involved in each of the 41,803 incidents in which two students were suspended and where I have prior year test scores for both students. Comparing the punishments of two students involved in the same incident is a reasonable strategy, because the majority of incidents (sixty percent) involving two students being suspended result in the students receiving different punishments. The fact that the assignment of different suspensions for the same incident is the norm, rather than the exception, lends additional credibility to the notion that schools may punish students differentially based on their potential contribution to the school's aggregate test performance. Indeed, I find this to be the case. While schools always tend to assign harsher punishments to low-performing students than to high-performing students throughout the year, this gap grows substantially during the testing

window. Moreover, this testing window-related gap is only observed for students in testing grades. In summary, schools apparent act on the incentive to re-shape the testing pool through selective discipline in response to accountability pressures.

Background: Student testing in Florida

Students in Florida have been tested on nationally norm-referenced tests, such as the Iowa Test of Basic Skills or the Comprehensive Test of Basic Skills, since as least the early 1990s, and in most cases prior to that. At the time period of this study, major changes occurred both in the nature of the tests administered and the public use of these tests for reporting and accountability purposes. In 1997, Florida first administered the FCAT, Florida's first test explicitly designed to align not to general national norms of basic skills but rather directly to the state's standards. Scores on the FCAT were reported beginning in 1998, and public pressure was immediately put on schools to attain high levels on the FCAT.

While today the FCAT is administered to students in every grade from three to ten, prior to 2001 it was only administered in grades four, eight and ten in reading and writing, and in grades five, eight and ten in mathematics. In the "off-grades" (and indeed, in the FCAT grades as well) a nationally norm-referenced test was administered as well. In 1998, the first year of public school reports on FCAT performance, schools faced no rewards or sanctions based on high or low performance on the FCAT; the accountability system in place at the time (which, indeed, had few explicit rewards or sanctions) was actually based on the norm-referenced tests, rather than on the FCAT, and the principal consequence of poor FCAT performance in 1998 was public ignominy.

Beginning in 1999, explicit rewards and sanctions were associated with FCAT performance, with very high-achieving schools (as well as schools improving considerably from one year to the next) receiving grants of $100 per student and very low-achieving schools receiving the label of "F" and the threat of vouchers for students in the future. (Prior to 1999, low-performing schools were labeled as "critically low-performing.") During the time period of the study, there is little evidence that principals or teachers faced dismissal as a result of low performance; the primary punishment for low performance was public stigmatization. Likewise, teacher and administrator pay were not linked to test performance during this time period.

In both accountability regimes, most schools were at the margin in some way: Prior to 1999 schools were divided into four groups, and beginning in 1999 schools were divided into five groups, with few schools in the top-rated group. Therefore, virtually all schools can be thought of as subject to the accountability regimes—particularly the 1999 regime. It is not obvious which schools should be more affected by the accountability systems: While those serving low-performing students are more susceptible to being stigmatized with low-performing labels, those serving high-performing students face yardstick competition from one another. The public was certainly paying attention to differences at the "top" of the market: Figlio and Lucas (2004) show that the housing market responded substantially to public reports of school grades among high-performing schools.

In sum, the accountability pressure faced by schools during the time period of this study was almost exclusively information pressure, rather than explicit rewards and sanctions. High-performing schools began receiving additional financial rewards in

1999, and low-performing schools faced the potential for future school vouchers beginning in 1999, so the incentives for elevated performance became somewhat greater across the spectrum beginning in 1999. However, I have no ex ante priors regarding which schools would respond most to accountability pressures by altering discipline. Low-performing schools may face a greater stigma threat associated with school accountability, but high-performing schools also face performance pressures, and it may be easier for high-performing schools to identify potentially low-performing students.[1]

Identification strategy and data

Students are punished for many reasons. However, the available data only identify the incident in which the involved student is involved and the length of suspension assigned to the student. To reduce the possibility that unobserved infraction severity might be driving the results, I limit my analysis to incidents with two suspended students and control for incident fixed effects. Therefore, I compare the attributes of the two students involved in the same incident, rather than comparing across incidents. Over the course of a given year, 10.7 percent of students are suspended at some point, and 3.5 percent of students are suspended at some point in a two-suspension incident. Nineteen percent of suspensions are for five or more days, and eight percent of suspensions are for ten or more school days.

While I cannot identify the school districts involved in this analysis, I can provide some basic information about the analysis sample itself. The analysis covers all two-

[1] When I repeat the analyses below for schools across the performance spectrum I find similar estimated results, implying that all types of schools responded similarly to accountability pressure along this margin. The strongest measured response was in the ten percent of most affluent schools, though this response was not statistically distinguishable from the responses of other schools.

student incidents (where both students were suspended from school) where the prior year's norm-referenced test score is observed for both students. In practice, then, all students in grades four through ten who were in the school district in the previous year are included in the analysis. Some second and third graders are included in the study, depending on the school district.[2] It is extremely unusual for kindergartners through third graders to be suspended, let alone be suspended in two-student incidents, so the lack of prior test scores at the early grades is irrelevant for this analysis. Not having eleventh or twelfth graders in the analysis is more troublesome, but since the highest grade tested is tenth grade, this absence should not influence the analysis either.[3]

There are 41,803 incidents in which two students with known prior test scores were both suspended. In total, 48,206 students participated in these incidents, indicating that 38 percent of those suspended participated in more than one multi-student suspension incident over the course of the study period. Seventeen percent participated in three or more multi-student suspension incidents over the study period, and two percent participated in more than five multi-student suspension incidents over the study period. Forty students (0.1 percent) participated in more than ten multi-student incidents over the four-year period. These incidents took place in 504 elementary, middle and high schools, though in practice the majority of incidents took place in the 104 middle and high schools where at least 100 multi-student incidents occurred. Fifteen percent of the students involved in multi-student incidents were elementary students (of these, more than two-thirds were fifth graders), and 63 percent of those observed were in middle school grades.

[2] In the districts covered in the analysis, students generally took norm-referenced exams in grades three through nine, and sometimes in grades one or two through nine.

[3] Below I mention that the results of the analysis reported in the paper are substantively unchanged if I include all suspended students regardless of whether I know their prior test scores, and include indicator variables for lack of a prior test score.

My identification strategy involves determining whether students of differing types receive different punishments over the testing cycle. In each two-person incident (for which both students were suspended) the students can receive either the same penalty or differential penalties. Because I am comparing between students who were suspended for the same incident, I only derive identification from the incidents generating differential penalties. I do not have information on either the nature of the incident or any cross-student assignment of blame; I only observe a unique identifier of each incident, and the dates of the suspensions administered.

In nearly sixty percent of cases, two students suspended for the same incident receive differential suspensions. Table 1 presents some descriptive information concerning these suspensions. One observes that on average, a suspension lasts for 1.97 days, with 19 percent of suspensions lasting one week or more and 8 percent of suspensions lasting two weeks or longer. These patterns are extremely similar for two-suspension incidents, suggesting that two-student incidents are not unusual, at least in terms of average suspension lengths. Moreover, one observes that, across the school year, some students tend to receive longer suspensions than others do: For instance, students predicted to score in the lowest proficiency group in reading and mathematics on the FCAT average suspension durations of 2.35 days with 23 percent receiving suspensions of one week or longer, while other students average 1.91-day suspensions, with 18 percent receiving one-week or longer suspensions. Black students average considerably longer punishments than do white students, low-income students average longer suspensions than more affluent students, and males tend to receive longer suspensions than females. It may be that low-scorers, low-income students, black

students and male students commit worse offenses or are more culpable than are other students, but it may also be the case that schools tend to discriminate against some groups of students. The purpose of the present paper is to determine whether these patterns vary across the testing cycle and between high-stakes and low-stakes grades.

In general, students in high-stakes grades receive slightly shorter suspensions than do students in low-stakes grades, and students in general receive similar suspension durations during the testing window (in January or March, depending on the year) as during the rest of the year. My identification of the effects of high-stakes testing on school suspension activity focuses on the interaction between the testing calendar, the grade level of the student, and the expected performance level of the student:

$$\begin{aligned}(\text{Suspension duration})_{sit} = {} & \gamma_{it} + \varphi(\text{Testing window})_{it} + \phi(\text{High-stakes})_{st} \\ & + \eta(\text{Low-scorer})_{st} + \kappa(\text{Testing window})_{it}(\text{High-stakes})_{st} \\ & + \varsigma(\text{Low-scorer})_{st}(\text{High-stakes})_{st} + \mu(\text{Low-scorer})_{st}(\text{Testing window})_{it} \\ & + \alpha(\text{Low-scorer})_{st}(\text{High-stakes})_{st}(\text{Testing window})_{it} + \theta X_{sit}\end{aligned}$$

for student s involved in suspendable incident i in time t. The testing window coefficient measured herein is in practice subsumed into the incident-specific fixed effect, rather than separately estimated, because whether or not the incident overlaps the testing window is invariant for any given incident. I measure the testing window as the period in which the testing takes place, as well as the week immediately preceding the test's administration. The vector X includes several variables: First, because of the potential that one student involved in the incident may be a recidivist and the other may be a first-time offender, I control for an indicator for whether the student in question is a first-time suspendee. In addition, to account for the potential that behavior and suspension patterns differ systematically over the school year (overall, as well as for different types of

children) I control for a set of month-of-year dummy variables and interactions between month-of-year dummies and the low-scorer variable.

The coefficient α is the key parameter of interest, and reflects the differential magnitude of the prior score-suspension gradient for students in high-stakes grades relative to low-stakes grades during the testing window versus other times of the year. Alternative measures of the dependent variable include indicators for whether the student is suspended for five or more days, and whether the suspension covers ten or more days.

I exploit several forms of variation: First, I take advantage of variation *across the school year* by estimating different relationships between a student's prior test score and his or her punishment during the testing window and the remainder of the school year. It is possible that low-achieving students are more likely to be more culpable during the testing period, but later in this paper I present evidence to indicate that this is unlikely to be driving my results. I also take advantage of variation *across grade* by estimating different testing window deviations in punishment for students in high-stakes grades versus those in low-stakes grades. In the low-stakes grades, schools have no incentive to differentially punish potentially low-achieving students. It should be noted that students in nearly all grades take standardized tests that could play a role in course placement and grade promotion during this testing window, so while the tests have roughly the same stakes attached for students across grades, they have different stakes for the schools. Furthermore, I estimate separate analyses *across time* to observe whether the patterns observed after 1997, when the tests had higher stakes for the school, were present in the 1996-97 school year, when the test scores were not reported.

I identify likely low scorers because in the years covered in this paper, the state consistently focused on the fraction of students who attained basic proficiency levels, rather than distinguishing among higher levels of proficiency. The prior test score variable used to predict low FCAT performance is the student's combined reading and mathematics score from the previous year on either the Iowa Test of Basic Skills or the Stanford-8 examination, depending on the school district, divided by the grade-year average of that test, to facilitate cross-time, cross-test, and cross-grade comparisons. In practice, I divide the prior test score distribution into two basic groups—students predicted to score level 1 (below basic proficiency) in the reading and mathematics portions of the FCAT, and all other students. I constructed these groups by regressing each student's realized FCAT scores on his or her immediate prior year test scores. I then identified a threshold prior test score that would predict low performance on the FCAT; I chose a threshold such that at that prior score 35 percent of students would be predicted to score level 1 on the FCAT. The results reported herein are not sensitive to this distinction. I experimented with thresholds ranging from 25 percent predicted to score level 1 to 50 percent predicted to score level 1 and the fundamental results of the paper were unchanged. Because students with disciplinary problems tend to be lower-achieving in general, 42 percent of suspended students are predicted to score level 1 on the FCAT, a rate almost fifty percent higher than the general student population.

The results reported in this paper are also unaffected by my decision to identify predicted low scorers rather than to explicitly include prior test scores. In prior versions of this paper I reported estimates of the suspension-prior test score gradient in which prior test scores were continuous rather than discrete, and consistently found that the

lower the prior test score, the more likely a student was to be differentially suspended during the testing window. These results are more difficult to conveniently interpret in the difference-in-difference-in-difference analysis, so I chose to report the more easily interpretable results for the paper. However, all results are highly consistent with one another and are available on request.

Prior test scores were unavailable for 14 percent of suspended students. I restrict my analysis only to incidents with two suspended students because I must directly compare suspensions for the two students based on their prior test outcomes. I repeated the analysis with all students in the regression, but with interaction terms with an indicator variable for whether prior test scores were unknown, and found that schools apparently treat students without prior test scores comparably to how they treat non-low-achievers, and that the presence of these interactions does not influence the findings reported in the paper.

Selection problems

The success of the identification strategy depends on whether several aspects of student misbehavior and punishment are true. First, it must be the case that low-achievers (measured by students predicted to score level 1 in reading and mathematics on the FCAT) tend to get into trouble at similar rates, relative to high-achievers, during the testing window as in other times of the year. The available evidence suggests that this is true. Low-achievers in high-stakes grades are 2.60 times as likely as high-achievers to be suspended—regardless of how many children are involved--during the testing window, but are 2.57 times as likely during the rest of the year. Low-achievers in low-stakes

grades are also 2.60 times as likely to be suspended during the testing window, and 2.61 times as likely during the rest of the year. These patterns are almost the same when limited to two-student incidents, the backbone of this project's identification problem: Conditional on being in a two-student incident, low-achievers in high-stakes grades are 2.53 times as likely as high-achievers to be suspended during the testing window, but are 2.54 times as likely during the rest of the year. Low-achievers in low-stakes grades in two-student incidents are 2.61 times as likely to be suspended during the testing window, and 2.58 times as likely during the rest of the year.

The fact that the ratios in student misbehavior rates are virtually identical across grades and at different times of the year indicates that the composition of misbehavers is not changing along with the testing cycle. Moreover, since these patterns are similar for all incidents and for two-person incidents, these results suggest that my identification strategy of comparing suspensions across two-person incidents is not introducing additional selection bias. Therefore, the potentially confounding selection problem associated with low-achievers intentionally misbehaving during the testing window in order to reduce their likelihood of taking the examination (or for any other reason) appears not to be substantial in this application.

It may still be the case that even though low-achievers do not differentially get into trouble during the testing window, as compared with high-achievers, they may still be more likely to be at fault in their suspensions than are high-achievers during this testing window. But the patterns of single-student incidents over the testing cycle closely mirror the patterns of two-student incidents mentioned above, and are also nearly identical for high-stakes grades and low-stakes grades. While I cannot tell whether these

incidents are differentially *severe* during testing periods, the fact that they are not differentially *numerous* increases confidence that a finding of differential suspensions during high-stakes testing periods is not being driven by student behavioral differences.

The identification strategy could also be confounded if parents of high-achieving students suspended during the testing window lobby principals for shorter suspensions to maximize the chances that their children take the test. But here too the evidence does not bear the selection story out: Conditional on suspension duration, there is no evidence that higher-achieving children suspended during the testing window ultimately took the test at higher rates than did low-achieving children. This finding suggests that a "taste for test-taking" among high-achieving families is not a major determining factor explaining the results of this paper.

Which students receive harsher suspensions during the testing window?

Table 2 presents estimates of the coefficients of the equation described above (except for the month dummies and the interactions between month dummies and prior test scores.) The data are for the first three years that the FCAT scores were reported (1997-98 through 1999-2000.) Each column represents a different version of the dependent variable: length of suspension, or a dummy variable reflecting a suspension of five or more days, or of ten or more days. One observes that across model specification, the coefficient on first-time suspendee is negative and statistically significant: First-time suspendees tend to receive average suspensions of one-third of a day shorter duration, are four percentage points less likely to receive a suspension of five or more days, and are two percentage points less likely to receive a suspension of ten or more days. The

interaction between high-stakes grades and the testing window is also negative and statistically significant, indicating that schools tend to be more lenient *in general* during the testing window; this result is expected, because the Florida Department of Education investigates schools that have large numbers of schools missing the FCAT examination, and in the later years of this analysis, schools that did not have a sufficiently high FCAT test-taking rate were sanctioned by the state.

The coefficient of interest for the present paper, however, is the three-way interaction between high-stakes grade, testing window, and predicted low achievement. Across the three specifications of the dependent variable, this coefficient is positive and statistically significant at conventional levels, indicating that while schools reduced their suspension penalties for higher-achievers in high-stakes grades (grades four, five, eight and ten) during the testing window, they raised their suspension penalties for lower-achievers in these same grades at this time. Moreover, the magnitudes of the coefficient estimates are striking—they suggest that the differential in the prior score-suspension gradient during the testing window in high-stakes grades versus low-stakes grades is between 1.7 and 2.2 times the magnitude of the coefficient on first suspension.

I also have data for the 1996-97 school year, when the FCAT test was administered but its results were not publicly reported. The very last row of Table 2 presents fixed-effects estimates of the differential suspension behavior during the year in which the FCAT test was administered in the high-stakes grades, but no grades really had high stakes attached to the test. One observes that the estimated effect of testing on differential suspensions of low-achievers and higher-achievers is not present in the year prior to the introduction of high stakes for schools, and in fact, the point estimates,

though not statistically significant, are of the opposite sign of those found in the years in which the FCAT was publicly-reported. The difference-in-difference-in-difference-in-differences between 1996-97 and 1997-98 through 1999-2000, 0.615 days longer suspensions, 13.3 percentage points increased likelihood of being suspended for five or more days, and 12.2 percentage points increased likelihood of being suspended for ten or more days, are all statistically significant at the ten percent level. The results are stronger still if I compare 1996-97 to 1998-99 and 1999-2000, the years of the strengthened "A-Plus" accountability regime under Governor Jeb Bush; these results have magnitudes of about 25 percent higher than those reported in the table, and are more strongly statistically significant than those reported in the table. I present the more conservative results in the table.

Might these relationships merely reflect some possible alternative pattern? While the estimated differences in suspension patterns over the testing cycle are strongly consistent with the notion of selective punishment, it may be that low previous performance is merely an indicator for some other unmeasured student attribute. To gauge the believability of this argument, I also control for, directly as well as with all two-way and three-way interactions, indicators for whether the student is black, free lunch eligible, or male. Table 3 presents the difference-in-difference-in-difference results of models that control for all of these variables and interactions as well. (All main effects and two-way interactions are included in the model, but are omitted from the table for purposes of parsimonious presentation.) As can be seen, the results are virtually unaffected by the inclusion of controls for race, socio-economic status, and sex, and their interactions with grade, the testing window, and the test window-grade interaction.

Therefore, if the results reported above truly reflect the effect of some unmeasured student attribute, that attribute must be correlated with prior performance but not highly correlated with student race, family income, or sex.

Do suspended students miss the test?

This analysis focuses on suspension durations because the identification strategy involves comparing students' punishments at one time of the year (the testing window) to other periods during the year. That said, one might be concerned that the differential suspensions during the testing window do not actually differentially affect the likelihood of a student ultimately taking the FCAT test. If students suspended during the testing window ultimately end up taking the test, then schools would face less incentive to suspend likely low-achieving students (though they would still have an incentive to do so if these students were considered more likely to be disruptive during the regular testing period.) It is possible for students to miss the FCAT due to suspension because, while there are scheduled make-up periods during the testing window, there is not an unlimited number of opportunities to make up the exam, and the make-up test must be completed during the assigned window.

While it is impossible to estimate a set of parallel regressions in which the dependent variable is suspended and missed the exam (since this would not be measured during the non-testing window portions of the year), it is possible to shed some light on the subject by comparing whether, among the two students suspended for the same incident during the testing window, the student predicted to score poorly on the FCAT is more likely to ultimately miss the examination. I find that, holding constant incident

fixed effects, among the two students suspended for the same incident during the testing window, a student expected to perform poorly on the FCAT is 12.3 percentage points less likely to take the FCAT than is a student expected to perform well. While small sample sizes reduce the precision of this estimate, it is statistically significant at the 14 percent level. As a falsification exercise, I also look at whether low-performing students suspended at times *other than the testing window* were less likely to ultimately take the FCAT; I find that they were only 0.1 percentage points less likely to take the FCAT than were their higher-performing co-suspendees. These results provide suggestive evidence that students with long suspension durations during the testing windows were indeed less likely to take the FCAT.

Suspension behavior and school test outcomes

I next explore whether schools that differentially suspend students end up with improved outcomes on the high-stakes exam. Here, I estimate a school fixed effects model in which I regress either the FCAT reading or mathematics examination score or an indicator for whether the student scores level 2 or above on the FCAT examination on a measure of the differential treatment of students during the testing window. I measure the school's suspension behavior in a given year as

$$\text{Suspension differential} = (LL_{window})/(LH_{window}) - (LL_{other})/(LH_{other})$$

where LL is the average length of a suspension for a low-achiever and LH is the average length of a suspension for a high-achiever. The subscript window reflects the testing window and the subscript other reflects the other times of the year. This suspension differential variable is measured using only students in high-stakes testing grades, and is

highly correlated with similar measures using the other two variants of the dependent variable. I measure this variable for each school and year, and follow the same schools over time to see if schools that give differentially long suspensions to low-achievers during the testing window in a certain year have students with particularly good test scores that year. As before, robust standard errors are adjusted for within-school, within-time clustering of standard errors. I only have access to FCAT scores from 1997-98 through 1999-2000.

The first column of Table 4 presents the results of this exercise. One observes that a one-standard-deviation increase in this measure is associated with a 2.3 point increase in the average FCAT reading scale score and a 3.8 point increase in the average FCAT mathematics scale score. While both of these estimated relationships are statistically significant, they are rather small in magnitude, given that the standard deviation in FCAT test scores is over 40 points. But as noted above, the largest degree of test window manipulation involves the margin between students predicted to score level 1 on the FCAT and those predicted to score level 2 on the FCAT. The second column of Table 4, therefore, concerns whether the student in question has attained level 2 or better. One observes that a one-standard-deviation increase in the test window manipulation measure is associated with a 1.2 percentage point increase in the likelihood that a student will attain level 2 or better on the FCAT reading examination and a 1.7 percentage point increase in the likelihood that a student will attain level 2 on the FCAT mathematics examination. Given that around one-third of students attained level 1 on the FCAT, this estimated effect, while still rather modest, may make a difference for schools on the margin. It should be noted, however, that these performance effects are likely upper

bounds, given that schools that are behaving along these margins may also be attempting to influence test scores along a variety of other measures as well. Nonetheless, the performance effects estimated herein suggest that the results described above are behavioral.

Conclusion

This paper presents evidence that schools respond to high-stakes testing by selectively disciplining their students. Schools have an incentive to keep high-performing students in school and low-performing students out of school during the testing window in order to maximize aggregate test scores. The evidence is supportive of this hypothesis—these patterns are precisely what are observed in the data, **but only for students in grades that are tested with high stakes for the school.** Since students suspended during the testing window are significantly more likely to miss the examination, this result suggests that schools may be deliberately attempting to reshape the testing pool in response to high-stakes testing. This finding is not observed in the time prior to the introduction of high stakes associated with the testing. These results indicate that schools may be using student discipline as a tool to manipulate aggregate test scores.

These results have significant implications for the design and implementation of school accountability systems. Accountability systems, no matter how well-designed, will have many incentives embedded within them for schools to "game the system." The successful design of accountability system hinges on the identification and closure of as many of these loopholes as possible. However, the likelihood that schools will find other

mechanisms through which they can inflate their observed test performance for the purposes of accountability suggests that all aggregate test scores should be taken with a grain of salt, and not viewed as perfect indicators of school productivity. Other indicators of school productivity, such as gain scores, that are harder to "game" may provide fewer incentives for schools to influence test scores through methods other than bona fide school improvement.

The results also may have effects beyond the testing system. Jacob and Lefgren (2003) show that children who are home from school are more likely to engage in criminal activity. If lower-achievers are disproportionately predisposed to committing criminal acts, then it is possible that some forms of high-stakes testing may also influence rates of criminal behavior. This last point, however, is only speculation.

References

Cullen, Julie and Randall Reback. "Tinkering toward accolades: School gaming under a performance accountability system." University of Michigan working paper, 2002.

Deere, Donald and Wayne Strayer. "Putting schools to the test: School accountability, incentives and behavior." Texas A&M University working paper, 2001.

Figlio, David and Lawrence Getzler. "Accountability, ability and disability: Gaming the system?" National Bureau of Economic Research working paper 9307, 2002.

Figlio, David and Maurice Lucas. "What's in a grade? School report cards and the housing market." American economic review, 2004.

Figlio, David and Joshua Winicki. "Food for thought: The effects of school accountability on school nutrition." Journal of public economics, forthcoming.

Jacob, Brian. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools." Journal of public economics, forthcoming.

Jacob, Brian and Lars Lefgren. "Are idle hands the devil's workshop? Incapacitation, concentration and juvenile crime." American economic review, 2003.

Skiba, Russ and Reece Peterson. "The dark side of zero tolerance: Can punishment lead to safe schools?" Kappan, January 1999.

Table 1: Suspension durations, by student type

Student type	Mean suspension duration	Fraction suspended for 5+ days	Fraction suspended for 10+ days
Full sample	1.97	0.19	0.08
Two-suspension incidents	2.06	0.20	0.08
Students predicted to be low-achieving	2.35	0.23	0.10
Students predicted to be higher-achieving	1.91	0.18	0.08
Black students	2.33	0.23	0.10
White students	1.68	0.15	0.07
Free lunch eligible students	2.05	0.20	0.08
Students not eligible for free lunch	1.80	0.18	0.08
Male students	2.07	0.20	0.08
Female students	1.76	0.17	0.07
High-stakes testing grade	1.87	0.18	0.08
Other grades	2.04	0.20	0.08
During testing window	1.94	0.19	0.08
Other times of the year	1.98	0.19	0.08

Table 2: Differential suspension patterns over the testing cycle, 1997-98 through 1999-00
Comparing students in two-student incidents; incident fixed effects models

Variable	Dependent variable		
	Suspension duration	Suspended for 5 or more days	Suspended for 10 or more days
First suspension	-0.347	-0.039	-0.019
	(0.042)	(0.005)	(0.004)
Low achieving student	0.469	0.037	-0.002
	(0.385)	(0.050)	(0.036)
High-stakes grade	-0.097	-0.015	-0.003
	(0.054)	(0.007)	(0.005)
Low achieving student x Testing window	0.050	0.028	0.014
	(0.362)	(0.047)	(0.034)
Low achieving student x High-stakes grade	-0.191	-0.024	-0.011
	(0.128)	(0.017)	(0.012)
Testing window x High-stakes grade	-0.239	-0.025	-0.016
	(0.116)	(0.016)	(0.010)
Testing window x High-stakes grade x Low achieving student	0.604	0.071	0.041
	(0.277)	(0.037)	(0.025)
Testing window x High-stakes grade x Low achieving student in 1996-97 [prior to FCAT reporting]	-0.011	-0.062	-0.081
	(0.569)	(0.094)	(0.083)

Notes: Robust standard errors adjusted for clustering are in parentheses beneath coefficient estimates. All models also control for month-of-year effects and interactions between month dummies and the low achieving student indicator. A "testing window" main effect is implied but subsumed within the incident fixed effect. High stakes grades are grades four, five, eight and ten. The testing window varies from year to year, but is always either in January, March, or both. Low achieving students are those who would be predicted to score level 1 on the FCAT reading and mathematics exams, given their prior year's performance on the Stanford-8 examination or Iowa Test of Basic Skills.

Table 3: Difference-in-Difference-in-Difference coefficients:
Including race, free lunch, and sex interactions
Comparing students in two-student incidents; incident fixed effects models

Interaction	Dependent variable		
	Suspension duration	Suspended for 5 or more days	Suspended for 10 or more days
Testing window x High-stakes grade x Low achieving student	0.618 (0.279)	0.068 (0.037)	0.046 (0.026)
Testing window x High-stakes grade x Black student	-0.047 (0.092)	-0.013 (0.012)	-0.009 (0.008)
Testing window x High-stakes grade x Free lunch-eligible student	-0.018 (0.095)	-0.014 (0.013)	0.006 (0.009)
Testing window x High-stakes grade x Male student	0.120 (0.096)	0.010 (0.013)	-0.007 (0.008)

Notes: Robust standard errors adjusted for clustering are in parentheses beneath coefficient estimates. All models also control for month-of-year effects and interactions between month dummies and the low achieving student indicator. In addition, all models control for testing window, low-achieving student status, race, sex and free lunch status, and testing window, as well as two-way interactions between testing window and high-stakes grade, as well as testing window or high-stakes grade and low-achieving student, black student, free lunch-eligible student, and male student indicators. A "testing window" main effect is implied but subsumed within the incident fixed effect. High stakes grades are grades four, five, eight and ten. The testing window varies from year to year, but is always either in January, March, or both. Low achieving students are those who would be predicted to score level 1 on the FCAT reading and mathematics exams, given their prior year's performance on the Stanford-8 examination or Iowa Test of Basic Skills.

Table 4: Estimated relationship between suspension differentials
and school FCAT performance, 1997-98 to 1999-00:
School fixed effects models,
effect of one-standard deviation increase in manipulation measure

Test	Dependent variable: FCAT scale score points	Dependent variable: Score level 2 or above on FCAT
FCAT reading test	2.329	0.012
	(0.969)	(0.006)
FCAT mathematics test	3.834	0.017
	(1.854)	(0.009)

Notes: Robust standard errors adjusted for clustering are in parentheses beneath coefficient estimates.