Testing, Crime and Punishment

David N. Figlio
University of Florida and NBER

Draft: January 2003

I am grateful to two undisclosed counties for providing me with the data to conduct this project, and to the National Science Foundation for research support. I benefited from conversations with Eric Brunner, Julie Cullen, Jens Ludwig, Steve Raphael, Lorien Rice, and participants at the AEA and APPAM meetings and seminars at the Universities of Chicago and Florida. I alone am responsible for errors.

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 (2002), have shown that schools tend to respond to accountability systems and testing regimes by classifying more marginal students as disabled. 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.

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. I contend that discipline can assist in this endeavor in at least two ways. First, 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. Empirical evidence suggests that students suspended during the testing window are 2.2 percentage points less likely to take the mathematics examination and 2.6 percentage points less likely to take the reading examination than are students suspended either before or after the testing window, implying that discipline outcomes could influence the testing pool. Given that the overwhelming majority of eligible students ultimately take the examination, suspended students comprise a large share of the students who do not take the test.

On the other hand, during the period immediately prior to the testing window (the so-called "cram period" where, according to anecdotal evidence, schools focus almost exclusively on test-preparation skills) schools may want to keep the more marginal students in school to maximize their test-preparation time. Schools have less of an incentive to keep in school students projected to perform at a very low level (and thereby likely to fail the examination regardless of test preparation) and indeed, may wish to punish these students more harshly to avoid their becoming a distraction during the preparation period. Likewise, schools have less of an incentive to treat likely very high-performing students leniently during this pre-testing period, because these students are likely to pass the examination with or without the test preparation practice (and indeed, they too may become a distraction to the preparation process.) Schools across the performance level distribution have the incentive to "game the system" 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 two populous counties in Florida. (For confidentiality reasons, these counties must remain unidentified in this research.) 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, Florida's high-stakes examination used to evaluate schools. 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 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. 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 grown substantially during the testing window. Moreover, this testing window-related gap is only observed for students in testing grades. In addition, the results suggest that during the cram period prior to the testing window, schools tend to assign the shortest suspensions to students in the middle of the prior year's test score distribution, implying that marginal students likely receive the most lenient suspensions in this case. Again, this result is only found for students in high-stakes testing grades. I also find that, conditional on incident fixed effects, the student receiving the harsher punishment is more likely to be suspended again than is the student receiving the lighter punishment.

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.

To analyze whether schools differentially suspend low-performing students during the testing window, I regress the student's suspension duration against an indicator for whether the incident occurred during the testing window and the interaction between a testing window indicator and the student's prior year test score. I measure the student's prior year test score as the sum of the student's prior year scores on the Stanford 9 mathematics and reading examinations. To correct for the fact that Stanford 9 scores increase with grade level on average, I standardize each student's score such that the mean prior test score for each grade level is zero and the standard deviation is one. (I also control directly for the student's pri or year test score.) In addition, I control for an indicator of whether the present incident is the student's first suspension of the year. I estimate this model separately for students in FCAT testing grades and non-testing grades, because the incentives to differentially suspend low-performing students during testing windows should be greater for students in testing grades.

To determine whether schools test to suspend marginal students more lightly during the test preparation period than during the time prior to the test preparation period, I estimate a similar incident fixed effects model using data from prior to the testing window. Here, I regress suspension duration against an indicator for whether the incident occurred during the preparation period, as well as interactions between this "cram period" indicator and the prior test score as well as the prior test score squared. (As before, I also control for the prior test score and the prior test score squared, as well as a first suspension indicator.) The quadratic term is important to capture the predicted U-shaped relationship between prior test score and suspension duration during the cram period for students in testing grades.

These identification strategies rely on students being given different suspensions for the same incident. As mentioned above, in sixty percent of the incidents involving two suspended students the students receive different penalties. The data suggest that students with different attributes systematically receive different penalties. Table 1 compares the average attributes of students who receive the more lenient penalty to students who receive the harsher penalty for the same incident. One observes that while students receiving lenient penalties have average test scores of one percent of one standard deviation below the mean for their grade level, students receiving harsher penalties have average test scores of 14 percent of a standard deviation below the mean. Conditional on the incident, students receiving harsher penalties are more likely to be African-American, free lunch eligible, and male. All of these differences are statistically significant at the one percent level. It is impossible to ascertain whether the two students merit penalties of different severity of varying culpability in the same incident, so one cannot know whether low-performing, low-income, male minority students systematically are more blameworthy in the incidents that also involve students with different attributes being suspended. However, it is harder to believe that these systematic differences in blameworthiness differ over the testing cycle, and differ across grade levels, so my identification strategy seems reasonable regardless of whether or not one believes that students with certain attributes are more likely to bear primary responsibility for the incidents in which they are involved. (***Address this head-on: do certain types of students get suspended more during the testing window? Might be test avoidance?***) The following sections provide empirical evidence on the differential punishments of low-performing students over the testing cycle.

Which students receive harsher suspensions during the testing window?

The first two columns of Table 2 present regression results on the differential suspension of low-performing students during the testing window. We observe that low-performing students tend to receive harsher suspensions over the entire academic year, regardless of whether they are in FCAT-tested grades. However, while there is no evidence that low-performing students in non-tested grades receive especially harsh relative punishments during the testing window, there exists statistically significant evidence of this pattern for students in grades covered by the FCAT examination. Specifically, the results suggest that for every standard deviation reduction in prior test scores, students in FCAT testing grades are suspended for an extra 0.19 days during the non-testing period and an extra 0.34 days during the FCAT testing window. Given that the typical suspension is just over two days in duration, this is a large effect, and the difference across the testing cycle is statistically significant at the six percent level. In contrast, for every standard deviation reduction in prior test scores, students in non-FCAT testing grades are suspended for an extra 0.23 days during the non-testing period and an extra 0.20 days during the testing window. Students in FCAT tested grades generally tend to have lighter suspensions during the testing window than do students in non-tested grades, a possible consequence of Florida's accountability requirements that sanction schools with low testing rates. Given that in 1999, the year for which I could match students to FCAT records in this dataset, students suspended during the testing window are 2.2 percentage points less likely to take the mathematics FCAT examination and 2.6 percentage points less likely to take the reading FCAT examination than are students suspended during the periods before and after the examination, these results suggest that schools may be attempting to substitute high performers for low performers in the testing pool. Interestingly, there is no relationship between whether the student has previously been suspended and his or her punishment, conditional on the incident.

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 replace prior year test scores with, in turn, each of three other students that in cross-section are correlated with differentially harsh student suspensions—indicators for whether the student is African-American (specification 2), free lunch eligible (specification 3) or male (specification 4). (Prior test scores are significantly negatively correlated with each of these indicators.) While these regression results suggest that African-American, low income and male students tend to be given harsher punishments conditional on the incident, one observes no evidence to suggest that this is particularly the case during the testing window for students in FCAT-tested grades. 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 correlated with student race, family income, or sex.

The final specification reported in Table 2 includes interactions between the testing window and each of these four student attributes (including prior test scores.) One observes that, even when controlling for the potential differential treatment of males, African-Americans, and low-income students over the testing cycle, the result described above—that low-performing students are differentially suspended during the testing window if their test scores would count against the school in FCAT aggregates—remains the same in magnitude and roughly the same in terms of statistical significance.

The point estimates reported in Table 2 actually imply that, on average, suspensions tend to be *lighter* during the testing window than at other times during the school year. School administrators face an extremely strong incentive to ensure that the overwhelming majority of eligible students take the examination; both accountability systems in Florida punished schools if fewer than 95 percent of the test-eligible student population did not take the test. Therefore, during the testing window, schools have an incentive to mete out considerably lighter punishments to students likely to pass the examination, and relatively tougher punishments to students unlikely to pass the examination. Given the constraints imposed upon schools, it is unsurprising that schools might attempt to reshape the testing pool by helping to ensure that misbehaving marginal-to-high-achievers are present for the examination while excluding misbehaving low-achievers from the test pool.

Which students receive harsher suspensions during the test preparation period?

As described above, schools have an incentive to suspend students differently during the test preparation "cram" period relative to the other times prior to the FCAT examination's administration. Here, I describe the test preparation window as the two weeks immediately preceding the FCAT examination, which is the modal amount of time students spend in intensive test preparation immediately prior to the FCAT administration, according to a survey that I conducted of 60 randomly-selected teachers in the two counties covered by this study. Unlike during the testing window, when schools' in centive to suspend students monotonically decreases with prior test scores (for students whose current scores would count for the school,) during the "cram" period this incentive should be nonlinear with respect to prior test scores. Specifically, I expect that students predicted to be on the margin of passing the FCAT examination will receive the least punitive suspensions, conditional on incident, relative to students considered either highly likely or highly unlikely to pass the exam regardless of test preparation.

Table 3 presents the results of an analysis comparing suspension durations for different types of students during the "cram" period versus the remainder of the pre - testing period. As the table makes clear, for students in non-tested grades, there is no apparent relationship between prior test scores and suspension duration during the test preparation period, relative to the other pre-testing period. However, there is an apparent relationship between prior test scores and suspension duration during the test preparation period for students in FCAT-tested grades. The interaction between the test preparation period and prior test scores is negative (though only significant at the 44 percent level) and the interaction between the test preparation period and the square of prior test scores is positive and statistically significant at the seven percent level. This result suggests a U-shaped relationship between test scores and suspension duration during the test preparation window. Interpreting the coefficients, the results indicate that the lightest suspension penalty during the test preparation period is assigned to students with prior test scores that are around one-quarter of one standard deviation above the mean score among the students facing disciplinary action. This translates to a score at about the 40th percentile in the overall testing distribution. Given that about two-thirds of students passed the FCAT during the first year of the examination, students scoring at about this level tended to be right on the margin of success on the FCAT. Therefore, schools apparently respond to the testing cycle by assigning the lightest suspensions to high-achieving students during the testing window and to marginal students during the pre-testing preparation period, exactly as might be expected.

Do lighter suspensions induce recidivism?

The preceding evidence suggests that schools respond to the testing cycle by differentially applying discipline to their misbehaving students. Do students in turn respond to these penalties when they consider whether to misbehave in the future? To address this question, I estimate another incident fixed effect model in which I simply regress one of two measures of recidivism—either an indicator for whether the student is suspended again for something else the same year or an indicator for whether the student is suspended again for something else sometime before the end of the dataset—against the measure of suspension duration and the indicator for whether the suspension in question is the student's first suspension. One observes that, conditional on the incident in which the student was involved, students who receive lighter suspensions are more likely to commit another offense serious enough to get suspended in the future. In addition, the results indicate that students who had been suspended already prior to the offense in question are more likely to be suspended in the future as well. Taken together, these results suggest that students who commit suspendable offenses tend to commit them serially, and that light penalties, conditional on incident, appear to induce future suspendable behavior. These results therefore indicate that testing-related disciplinary actions may have consequences that transcend the specific incidents in question.

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. On the other hand, schools have an incentive to keep marginal students in school during the pre-testing preparation period, and have less of an incentive to keep very high or very low-performing students in school during this so-called "cram" period. Again, this pattern is observed in the data. Taken together, 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.

References

To be added.

Table 1: Attributes of students facing different penalties for the same incident

Student attribute	Students received same punishment (40% of incidents)	Students received different punishments	
		Student with more lenient punishment	Student with harsher punishment
Average standardized test score	0.077	-0.009	-0.140
Percent African-American	48.9%	48.4%	57.6%
Percent free lunch eligible	65.7%	74.0%	76.8%
Percent male	66.2%	68.4%	71.2%

Table 2: Suspension patterns over the testing cycle

Specification	1T	1N	2T	2N	3T	3N	4T	4N	5T	5N
Test grade	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO
Prior test score	-.191 (p=.00)	-.234 (p=.00)							-.115 (p=.00)	-.140 (p=.00)
African American			.628 (p=.00)	.841 (p=.00)					.584 (p=.00)	.769 (p=.00)
Free lunch eligible					.123 (p=.06)	.251 (p=.01)			.011 (p=.87)	.104 (p=.28)
Male							.304 (p=.00)	.148 (p=.14)	.320 (p=.00)	.1168 (p=.09)
First suspension	.079 (p=.75)	-.164 (p=.65)	.098 (p=.70)	-.154 (p=.67)	.083 (p=.74)	-.189 (p=.61)	.081 (p=.75)	-.191 (p=.60)	-.126 (p=.62)	-.003 (p=.99)
Testing window	-.146 (p=.08)	-.024 (p=.83)	-.163 (p=.16)	-.076 (p=.62)	-.262 (p=.09)	-.018 (p=.93)	-.195 (p=.17)	-.015 (p=.94)	-.247 (p=.24)	-.046 (p=.86)
Testing window x prior score	-.146 (p=.06)	.029 (p=.77)							-.143 (p=.08)	.049 (p=.63)
Testing window x African American			.034 (p=.83)	.149 (p=.47)					-.049 (p=.77)	.162 (p=.44)
Testing window x free lunch eligible					.142 (p=.41)	-.006 (p=.98)			.110 (p=.53)	-.008 (p=.99)
Testing window x male							.059 (p=.73)	-.005 (p=.98)	.071 (p=68)	-.045 (p=.83)

Notes: The dependent variable is the suspension duration for the student. All regressions control for incident fixed effects. Data are for 41,803 incidents with two suspended students. Prior test scores are standardized by grade.

Table 3: Suspension patters during the pre-testing "cram" period

Specification	6T	6N
Prior test score	-.236 (p=.00)	-.316 (p=.00)
Prior test score^2	.024 (p=.47)	.089 (p=.05)
First suspension	-.314 (p=.31)	-.156 (p=.71)
Cram period	-.043 (p=.85)	1.139 (p=.33)
Cram period x prior score	-.155 (p=.44)	.178 (p=.80)
Cram period x prior score^2	.221 (p=.07)	-.033 (p=.91)

Notes: The dependent variable is the suspension duration for the student. All regressions control for incident fixed effects. Data are for 41,803 incidents with two suspended students. Prior test scores are standardized by grade.