NBER WORKING PAPER SERIES

ACCOUNTABILITY, ABILITY AND DISABILITY:
GAMING THE SYSTEM

David N. Figlio
Lawrence S. Getzler

Working Paper 9307
http://www.nber.org/papers/w9307

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2002

We are grateful to Sheila Murray, Richard Rothstein, and Jim Wyckoff, as well as seminar participants at Iowa State University, the National Bureau of Economic Research, Stanford University, University of California-Davis, and University of Florida, and participants at the APPAM and AEFA meetings for helpful comments and suggestions, to the National Science Foundation for research funding, and to six undisclosed school districts for generously providing us with the confidential data employed in this study. Any remaining errors are our own. The opinions expressed herein are those of the authors and do not necessarily reflect the positions of their employers nor those of the National Bureau of Economic Research.

Accountabilty, Ability and Disability: Gaming the System
David N. Figlio
Lawrence S. Getzler
NBER Working Paper No. 9307
October 2002
JEL No. I2

ABSTRACT

The past several years have been marked by a general trend towards increased high-stakes testing for students and schools and test-based school accountability systems. There are many potential school responses to testing programs. This paper investigates the potential that schools respond by "gaming the system" through reshaping the test pool. Using student-level panel data from six large counties in Florida, we study whether the introduction of the Florida Comprehensive Assessment Test in 1996 led schools to reclassify students as disabled and therefore ineligible to contribute to the school's aggregate test scores. Employing student-level fixed effect models and a series of secular trends as controls, we find that schools tend to reclassify low income and previously low performing students as disabled at significantly higher rates following the introduction of the testing regime. Moreover, these behaviors are concentrated among the low income schools most likely to be on the margin of failing the state's accountability system.

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

Lawrence S. Getzler
Virginia Department of Planning
and Budget
Ninth Street Office Building
200 North Ninth Street
Richmond, VA 23219-3418

Introduction

Education is currently at the forefront of the nation's political agenda: everyone, regardless of political persuasion, wants to see an improvement in the performance of U.S. schools. This consensus ends abruptly, however, when it comes to determining how to effect such a change in performance. One popular approach is to increase the accountability of schools to the public, by assessing schools on the basis of improvements in students' performance on standardized examinations and by offering remedies, such as increased choice (either within the public sector or through vouchers for private schools), reconstitution, or closure, in the event of persistent identified failure of a school to improve. Accountability measures have been proposed or implemented in dozens of states and going forward will be required in all states.

On January 8, 2002, President Bush signed into law the reauthorization of the Elementary and Secondary Education Act, also known as the *No Child Left Behind Act of 2001* (NCLB). A centerpiece of this education reform involves implementing a system of school accountability. States must design systems of school report cards based on the fraction of students demonstrating proficiency in reading and mathematics. Under NCLB, if students do not make adequate yearly progress, schools and districts face consequences such as mandatory public school choice and the possibility of complete school restructuring, as well as the redirection of federal funds; states risk the loss of federal administrative dollars. Additionally, the classifications or grades formally assigned to schools may affect the attractiveness of the local area to potential and current residents and the perceptions of local officials by the public. Figlio and Lucas (2000) provide evidence that housing markets are highly responsive to introduction of government-provided school report cards. Thus, the grading of schools using student test data provides numerous incentives for schools to "game the system."

Schools may react to these incentives by increasing class time spent on subjects and topics that are emphasized in the accountability exams, while decreasing class time on subjects and topics either not in or not emphasized in the exams. It should be noted that this type of strategy may be perceived by policy-makers as precisely the desired response to the accountability system, rather than as a "gaming" of this system. Significant class time may also be taken on test-taking strategies. Schools may even be less inclined to discourage poorer students from dropping out. For example, a Virginia school district superintendent said that the state's accountability exam system "actually encourages higher dropout rates … It is actually to the school's advantage to drop slow learners and borderline students from the school, because they are usually poor test-takers." (Borja, 1999) In part because of the newness of school accountability systems, we know of few attempts to seriously quantify school responses to these incentives.[1]

Another potential reaction to the incentives created by accountability systems involves the classification of students into special education categories exempt from taking the tests used for school grading.[2] Schools could potentially improve their state-assigned grade or classification by taking their poorest performing students out of the testing pool by classifying them into the special education categories exempt from taking the tests.[3] Additionally, the schools could potentially improve their state-assigned grade or classification by refraining from classifying better-performing students into the special education categories exempt from taking the tests. The American Institutes

[1] Papers that discuss these types of incentives include Elmore et al. (1996), Goldhaber (2002), Ladd (2001) and Koretz (1996). However, these are not empirical studies of school responses to incentives.   A few recent academic papers describe school responses to incentives embedded within accountability systems, other than the response described in this paper. Figlio (2002) finds that the introduction of accountability exams in Florida has resulted in fewer and shorter disciplinary suspensions for poor-performing students during the "cram period" prior to accountability exam testing dates. Figlio and Winicki (2002) show that Virginia schools threatened with sanctions tend to alter their nutrition programs during testing periods and substantially increase nutrients clinically shown to boost short-term cognitive performance. Jacob (2002) and others present evidence that schools subject to accountability systems may respond by retaining marginal students.

[2] The NCLB Act will require special education participation, but for reasons mentioned in the Discussion section of this paper, incentives to game the system through the classification of students into special education categories will remain.

[3] States may have other incentives to over-classify students into special education categories. For example, Cullen (2001) found that fiscal incentives could explain nearly 40% of the growth in student disability rates in Texas.

for Research's (AIR) new national study on special education costs helps demonstrate the potential flexibility and opportunity that school decision makers have in determining which, if any, special education category to place students in. AIR finds very wide variation in costs and services within single special education categories. In fact they find less than ten percent of the variation in special education costs in carrying out Individualized Education Plans can be explained by the exceptionality categories in the federal/state indicator record (Chambers et al, 2002). This implies that there may be significant discretion in how to classify individuals with specifically identifiable needs.

In this paper we use highly detailed student-level data to examine whether the initiation of the Florida Comprehensive Assessment Test (FCAT) has affected Florida public schools' decisions on special education assignments. Using student-level fixed effects models, we find that following the introduction of the FCAT testing program low-performing students and students from low socio-economic backgrounds were significantly and substantively more likely to be reclassified into disability categories exempted from the accountability system. These differences persist even after controlling for a rich set of time trends in disability classification. We also find that high-poverty schools are significantly more likely to reclassify low-achieving students than are more affluent schools.

While ours is the only paper to apply student-level fixed effects models to this topic, we know of two other current working papers that describe similar issues. Jacob (2002), looking at the effects of test-based accountability in Chicago, shows that low-achieving students in struggling schools are the most likely to be placed in special education, a finding similar to ours. While Jacob does not estimate student fixed-effects models, he does control for prior achievement test scores and background characteristics. Cullen and Reback (2002), using aggregate data and a clever identification strategy, exploit the discontinuity in rewards in Texas's accountability system to show that schools respond to incentives to shape the test pool. These two papers, taken together with ours, present complementary evidence--in three states and with three very different identification strategies--that schools respond to the incentive to classify marginal students into special education.

High-stakes testing in Florida

Beginning in the 1996-97 school year, students in certain grades began to take the Florida Comprehensive Assessment Test in reading and mathematics for the purpose of evaluating schools' performance in fostering educational achievement.[4] The FCAT tests were designed to align closely with the Sunshine State Standards, a set of core knowledge that students in particular grades are expected to know. The tests are challenging, and are generally accepted to be among the more comprehensive state-level student assessments. These tests were initially used by the state to identify low-performing schools, and beginning in 1999 were used to grade schools on an explicit A through F scale, though this new grading regime was not fully known at the time of our last year of testing in the present analysis. Students in fourth, eighth, and tenth grades were tested in reading and writing, while students in fifth, eighth, and tenth grades were tested in mathematics. No major changes occurred to special education financing in Florida over this time period.

All regular education students are required to take the FCAT examinations, but students in only a small number of disability classifications are required to take the exam. Specifically, all speech or language impaired or hospital/homebound students are required to take the FCAT. But in all other disability categories (educable or trainable mentally handicapped, orthopedically impaired, deaf or hard of hearing, visually impaired, emotionally handicapped, specific learning disabled, profoundly mentally handicapped, dual-sensory impaired, autistic, severely emotionally disturbed, traumatic brain injured, or developmentally delayed) FCAT test participation is determined by

[4] Students had previously taken the Florida Writes! writing assessment.

school personnel and the student's parents in the student's Individualized Education Plan, and test scores of all students in these categories are exempted from school accountability programs. While some of these disability categories are clearly more mutable than others, it is certainly possible that marginal students may be classified (or de-classified) from some of the exempted categories as a result of the testing regime.

Identification strategy and data

We are interested in investigating the effects of the testing regime on disability classification probabilities. Due to the numerous potential omitted variables problems in this application, we utilize panel data and estimate models with *student-level fixed effects* to capture any time-invariant student-level variation in the probability of disability classification. Therefore, we draw our identification from students whose timing of classification switches coincides with the timing of the testing regime. Since some students may be classified in anticipation of testing policy changes and others may experience delays in testing-related classification changes, this strategy yields *conservative* estimates of the effect of testing on disability classification. Because of the possibility that different types of students have become more likely to be reclassified into special education over time, we also control for *linear time trends in disability classification*. We estimate different models in which we in turn assume that all students' classification probabilities trend together over time, and in even more highly parameterized models in which we allow linear trends in classification probabilities to vary across different types of students or schools. This strategy should also serve to generate conservative estimates of the effects of testing on disability classification, because some of the change in disability classification associated with the testing regime would almost surely be captured by a time trend. We also have estimated models that include both attribute-specific linear time trends and year effects (to capture any nonlinear time trend in overall classification patterns.)

In these models, we cannot estimate an overall testing effect, because the testing regime began at the same time for the entire sample. However, we can still estimate the coefficients on the interaction terms between the testing regime and student or school attributes. In each case, these estimated interaction terms are virtually identical to those reported in the paper; therefore, we do not report two sets of regression results, and instead report only the set of results where it is possible to estimate an overall testing effect.

Our data come directly from the student records of six large herein-unidentified county-level school districts, each among the one hundred largest school districts in the United States.[5] Students in these school districts are more likely to be urban and are somewhat more likely to be racial or ethnic minorities than would a cross-section of Florida in general, but are large and diverse enough to have vast quantities of students of all socio-economic backgrounds, and schools at all levels of the socio-demographic spectrum.

School districts in Florida have uniform reporting requirements, and students are merged over time based on social security number, and in the event of no match by social security number, by first name, sex, race, and birth date. Students who change school districts over the study period remain in the study provided they relocated to another district included in the project. For the period from 1991-92 through 1998-99, we follow every student in kindergarten through eighth grade for all six counties. School district records include free lunch status, grade, and disability status. In addition, in two of these counties, we observe the student's Stanford 9 standardized test score for nearly every student in each year from 1994-95 through 1998-99. (Counties vary from year to year in which students are tested. In one county, students were tested beginning in grade one in some years and grade two in other years; in the other county, students were tested beginning in grade two in some years and grade three in other years.) All told, our dataset consists of 4,334,284 student-

[5] Counties participating in this study wish to remain unidentified.

year observations. We observe student background characteristics in 4,171,752 cases, and prior year Stanford 9 test scores in 907,577 cases. (Note that we have substantially fewer observations on prior test scores not because of sample attrition—94 percent of students in the two relevant counties have test score data—but rather because we only have Stanford 9 test scores for two of the six counties, and then for a shorter time window.) Due to the likelihood of error correlation at the school level, we adjust all standard errors for heteroskedasticity and clustering at the school level.

Table 1 shows the changes in disability classification rates in our population over time. We observe that the overall rate of disability classification increased over the period covered by this study. At the beginning of the study period, 7.3 percent of students were classified as disabled in categories that would eventually be test-exempt. By the end of the period, however, this classification rate had increased to 10.8 percent. While more of the increase in disability classification generally occurred following the introduction of the testing regime, there is an apparent trend in classification occurring prior to the testing period, implying that our decision to control for time trends is a prudent one. (Of course, some of the pre-testing run-up in disability classification could be in anticipation of the introduction of the testing system.) Table 1 also presents these figures for free lunch eligible students (a proxy for likelihood of performing poorly on the FCAT examination) and those who are not free lunch eligible. In the case of the free lunch eligible, classification rates increased from 8.7 percent to 10.6 percent in the period prior to the introduction of the testing regime, while in the case of more affluent students, classification rates remained relatively stable in starting at 6.1 percent and ending at 6.2 percent. After the introduction of the testing regime, the test-excluded disability classification rates increase substantively for both groups.

The left panel of Table 2 describes the transitions into disability classification, by grade, before versus after the introduction of the testing regime. The vast majority of students enter special education during the elementary grades, but one might reasonably expect that if the increase in reclassification is occurring as a result of the testing regime as opposed to general trends toward increased classification that the third-to-fourth-grade transition would see the largest spike in classification following the introduction of high-stakes testing, as fourth grade is the first year of testing with consequences for schools. We observe that there is no statistically significant or economically meaningful change in classification transitions from grade-to-grade after versus before the testing program's introduction in any of the elementary school grade transitions, *except for the transition into fourth grade.* In this transition, we observe increased propensities for students to be reclassified into test-exempt special education categories following the introduction of the FCAT testing program. This difference is significant at the one percent level when standard errors are adjusted to account for clustering of errors within schools.

The right panel of Table 2 breaks these transitions out separately for free lunch eligible students and more affluent students. We observe that the post-FCAT increase in disability classification during the third-fourth-grade transition is entirely due to increases in classification of low-income students. On the other hand, at no other transition does the post-FCAT effect on reclassification ever approach statistical significance for either low-income or more affluent students. This provides some suggestive evidence that schools may be responding to incentives to reclassify certain students as disabled in order to reduce their contribution to aggregate measures of test performance. Of course, whether these effects are causal remains to be seen.

Regression results

Table 3 describes the estimated effects of the introduction of high-stakes testing on test-excludable disability classification. Specification 1 reports the estimated mean effects of the introduction of testing, in a model controlling for student-level and grade-level fixed effects, but no time trends. We observe that the introduction of the FCAT test is associated with an increase in the likelihood that a student will be classified as disabled by 5.6 percentage points. This estimated effect is statistically significant at any reasonable level; it is also economically significant, as 8.9 percent of the sample of students are identified as having a test-excludable disability, implying that the introduction of FCAT testing is associated with a more than 50 percent higher rate of disability classification in the six counties in question.

While schools have a financial incentive to classify students as disabled regardless of background, this incentive should be particularly strong for students whom the school views as at risk of performing poorly on the standardized examination. Given that low-income students tend to do more poorly on standardized examinations than do higher-income students, one proxy for this screen might be free lunch eligibility. Therefore, the second specification of Table 3 includes an interaction term between testing and free lunch eligibility. We observe that while post-FCAT, the classification of more affluent students increased by an estimated 3.6 percentage points, the estimated change in classification associated with the change in testing regime is again as great for free lunch eligible students. Specification 3 adds a time trend to the model; here, we observe that while the estimated effect of the testing regime for more affluent students falls considerably, the estimated difference in the effects for free-lunch and non-free-lunch students remains virtually the same, and is still statistically significant at any reasonable level of significance. Specification 4 controls for separate time trends for low-income and higher-income students, and again the results clearly indicate that low socio-economic-status students are most likely to be reclassified in response to the testing policy, even after controlling for a rich set of time trends.

Specifications 5 through 8 from Table 3 present the results from these same four regressions, but only for the two counties where we also have Stanford 9 test scores. We observe that while the results are the same as those reported above, in terms of being strongly statistically significant, the estimated magnitudes of the results, though still quite large, are more modest than in the six-county case. This suggests that the models that follow that look at testing effects by prior test scores rather than socio-economic status may also generate relatively conservative estimates of the responses to the testing regime. However, we have no way of knowing for certain whether this is true.

Specification 9 from Table 4 presents the results from the parallel model to Table 3's Specification 6. Here, all variables are interacted with the student's Stanford 9 mathematics test score from the prior year rather than with free lunch eligibility. As with Specification 6, this specification does not control for time trends. The drawback of this exercise is that, due to data limitations, we can only observe one pre-testing year of data. But we still observe results that yield similar conclusions as the free lunch interactions do: the lower last year's test performance, the more likely a student is to be classified as disabled. Specifications 10 and 11 repeat the same model, but in turn add a general time trend, then a time trend interacted with the prior year's mathematics test score. We see that in both of these specifications, schools tended to increase disability classification post-testing disproportionately for students who performed poorly on the prior year's test.

Specification 12 from Table 4 presents the identical model as Specification 11 (all fixed effects and past-performance-specific trends) but changes the dependent variable to look only at a very specific classification decision. In this model, students are included in this specification only if they are either classified as learning disabled or have another disability *that does not automatically exclude them from testing on the FCAT*. This model is extremely highly parameterized, and because of the fixed effects included in the model, identifies the effects of testing entirely on the basis of students whose classification switches between learning disabled, and therefore test-excluded, and non-excluded disabilities. Even in this specification, which we present as corroborative evidence, the results stay consistently strong in magnitude and statistical significance, indicating that schools are more likely to switch low-performers from a test-included to a test-excluded disability following the introduction of the testing regime. Because of the relatively small number of classification-switchers, however, the remainder of the paper focuses on disability classification more generally, rather than this very specific type of classification decision.

Specifications 13 and 14 from Table 4 report the results of models in which students are grouped by school type, with the notion that certain schools might be more sensitive to a school accountability system than are others. We identify schools as "high poverty" if the school has more than the district-wide median fraction of free lunch-eligible students. Specification 13 controls for separate time trends for high-poverty and low-poverty schools, while Specification 14 further controls for prior-test-score-specific separate time trends for high-poverty and low-poverty schools. We observe that high-poverty schools are significantly more likely to reclassify students than are their relatively low-poverty counterparts. As Specification 14 demonstrates, these results are particularly concentrated for previously low-performing students. In summary, schools that ex ante are likely to be more threatened by a test-based accountability system, because they have a larger fraction of students likely to perform poorly on the examination, tend to be more aggressive in reclassifying previously low-performing students as disabled in an apparent response to the introduction of the high-stakes testing program.

Discussion

We have estimated that the introduction of the high-stakes FCAT testing is associated with a dramatically higher rate of disability classification. We have also determined that the probability that a low-performing student or a student from a low socio-economic background would be reclassified into a disability category exempted from the accountability system increased significantly after the introduction of the high-stakes FCAT examinations. In addition, we found that high-poverty schools are significantly more likely to reclassify students than more affluent schools.

Altering decisions on special education classification for students reduces the accuracy of the grades or classifications given to schools based on the accountability exams and profoundly affects the students' individual educational experience. Reduced accuracy in the grades or classifications given to schools based on the accountability exams reduces the potential effectiveness of public policy based upon that data.

The incentive to place the students likely to perform worst on the state tests into special education classes may cause schools to place in special education students whom they believe would be better off in other classes. Since many states have laws that limit the number of students per special education teacher, the placement of those students into special education classes who otherwise would not have been so placed may require that students who would benefit more from special education be prevented from taking special education classes.

Also, the cost of providing special education far exceeds the cost of traditionally educating a student. According to a new study by the American Institutes for Research, the ratio of spending per special education student to spending per regular education student is 1.90 on average. (Chambers et al., 2002) Thus, funds could be inappropriately spent on special education for students who may be better off in less costly traditional classrooms; schools could potentially spend those funds more productively if the incentives to alter special education assignments did not exist.

The NCLB Act will require that students that are classified into special education categories participate and be counted. Specifically, under the NCLB Act, all students in each defined subgroup[6] must meet or exceed the state's proficient level of academic achievement by the end of the 2013-14 school year. The legislation specifies intermediate goals for meeting this objective.

[6] Students with disabilities are one of several defined subgroups.

These include each state establishing "statewide annual measurable objectives" that include a "single minimum percentage of students who are required to meet or exceed the proficient level on the academic assessments." These minimum percentages apply separately to each subgroup of students, but not all subgroups must make adequate yearly progress each year. The subgroups that do not meet or exceed the minimum percentage still must decrease their percentage of students that are below proficiency by 10 percent when compared with the preceding year.[7]

Despite the requirement under the NCLB Act that all subgroups, including students with disabilities, be included in the accountability testing system, incentives to game the system through special education classification will remain. First, NCLB does permit testing accommodations for students with disabilities. Accommodations, such as additional time, can potentially aid any student's performance, including those students without legitimate or clear-cut disabilities. Thus, the incentive to over-classify[8] low-performing students and students from low socio-economic backgrounds into special education remains. Also, since all subgroups, including students with disabilities, will be required to have the same minimum percentage of members meeting proficiency or at least decrease the percentage of non-proficient students by 10 percent annually, schools will have the incentive to place "ringers" in the students with disabilities category. In other words, since it will likely be particularly difficult to have the students with disabilities subgroup reach the minimum percentage, schools will have a strong incentive to add students to that category who are likely to achieve proficiency. For example, schools would likely improve their probability of attaining adequate yearly progress for all subgroups if they were to place relatively high-achieving

[7] Source: *The No Child Left Behind Act of 2001*.

[8] Some may be of the opinion that prior to the accountability exams not enough students were receiving special education. If this opinion is accurate, then perhaps this incentive results in some students being better off. Still, as described earlier in this paper, this will likely cause schools to place in special education at least some students who would be better off in other classes. And since many states have laws that limit the number of students per special education teacher, the placement of those students into special education classes who otherwise would not have been so placed may require that students who would benefit more from special education be prevented from taking special education classes.

students with mild dyslexia into the students with disabilities subgroup, who would not have otherwise been so classified.

References

Borja, Rhea R., “Comments: SOLs Raise Concern, Little Support / Change Tests, Use Multiple Criteria, Speakers Say,” Richmond Times-Dispatch, December 1, 1999.

Chambers, Jay, Tom Parrish, Jamie Shkolnik, and Maria Perez, “A Report on the 1999-2000 Special Education Expenditures Project,” special session presented at the annual research conference of the American Education Finance Association, Albuquerque, NM, March 2002.

Cullen, Julie Berry, “The Impact of Fiscal Incentives on Student Disability Rates,” *Journal of Public Economics*, article in press, 2001.

Cullen, Julie Berry and Randall Reback, “Tinkering Toward Accolades: School Gaming under a Performance Accountability System,” Working paper, University of Michigan, 2002.

Elmore, Richard F., Abelmann, Charles H., and Susan H. Fuhrman, “The New Accountability in State Education Reform: From Process to Performance,” pages 65-98 in Holding Schools Accountable, Helen F. Ladd, editor, The Brookings Institution, Washington, D.C., 1996.

Figlio, David N., “Testing, Crime and Punishment,” National Bureau of Economic Research working paper, 2002.

Figlio, David N. and F. Joshua Winicki, “Food for Thought? The Effects of School Accountability on School Nutrition,” National Bureau of Economic Research working paper, 2002.

Figlio, David N. and Maurice E. Lucas, “What’s in a Grade? School Report Cards and House Prices,” National Bureau of Economic Research working paper no. 8019, 2000.

Goldhaber, Dan, “The Reauthorization of the Elementary and Secondary Education Act (ESEA): What Might Go Wrong with the Accountability Measures of the ‘No Child Left Behind Act?,’” a policy memo for the Thomas B. Fordham Foundation conference “Will No Child Truly Be Left Behind?: The Challenge of Making This Law Work,” February 13, 2002.

Jacob, Brian A., “The Impact of High-Stakes Testing on Student Achievement: Evidence from Chicago,” Working paper, Harvard University, 2002.

Koretz, Daniel, “Using Student Assessments for Educational Accountability,” pages 171-196 in Improving America’s Schools: The Role of Incentives, Hanushek, Eric A., and Dale W. Jorgenson, editors, Washington, D.C., National Academy Press, 1996.

Ladd, Helen, “School-Based Educational Accountability Systems: The Promise and the Pitfalls,” *National Tax Journal*, 385-400, 2001.

The No Child Left Behind Act of 2002, Public Law 107-10, 107th Congress, 1st Session 2002.

Table 1: Over-time changes in test-excluded disability classification rates, six Florida counties

School year	Overall classification rate	Classification rate of free-lunch-eligible students	Classification rate of non-free-lunch-eligible students
1991-92	7.3%	8.7	6.1
1992-93	7.8	9.3	6.1
1993-94	8.1	9.5	6.5
1994-95	7.8	9.7	5.2
1995-96	8.8	10.6	6.2
INTRODUCTION OF TESTING REGIME			
1996-97	9.4	11.0	7.6
1997-98	9.6	11.8	7.1
1998-99	10.8	13.2	7.4

Table 2: Grade-to-grade transitions in test-exempt disability classification, before versus after testing regime introduction

	Percentage of students classified in a test-exempt category in the following grade								
Among students NOT classified as disabled in grade:	General population			Free lunch eligibles			Non-free lunch eligibles		
	Pre-FCAT	Post-FCAT	Robust p-value of difference	Pre-FCAT	Post-FCAT	Robust p-value of difference	Pre-FCAT	Post-FCAT	Robust p-value of difference
1	3.0%	3.0%	.699	3.5%	3.5%	.968	2.3%	2.3%	.945
2	3.2	3.2	.987	4.0	4.0	.817	2.2	2.3	.267
3	2.7	2.9	.007	3.3	3.8	.000	1.9	1.9	.276
4	2.0	2.0	.608	2.5	2.6	.466	1.3	1.3	.455

Table 3: Estimated effects of testing on disability placement, by socio-economic status
(robust standard errors in parentheses beneath coefficient estimates)

Specification:	1	2	3	4	5	6	7	8
Student fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Grade fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Standard errors adjusted for school level clustering	YES	YES	YES	YES	YES	YES	YES	YES
General time trend included	NO	NO	YES	YES	NO	NO	YES	YES
Separate trends for low-income and high-income students	NO	NO	NO	YES	NO	NO	NO	YES
Coefficient on testing	0.056 (0.001)	0.036 (0.001)	0.010 (0.001)	0.009 (0.001)	0.046 (0.001)	0.027 (0.001)	0.002 (0.001)	0.012 (0.001)
Coefficient on testing x free lunch eligible		0.038 (0.002)	0.039 (0.002)	0.039 (0.002)		0.034 (0.002)	0.034 (0.002)	0.016 (0.002)
Number of counties	6	6	6	6	2	2	2	2

Table 4: Estimated effects of testing on disability placement, by prior mathematics test performance (robust standard errors in parentheses beneath coefficient estimates)

Specification:	9	10	11	12	13	14
Sample:	Full population	Full population	Full population	Learning disabled or test-included disabled students	Full population	Full population
Student fixed effects	YES	YES	YES	YES	YES	YES
Grade fixed effects	YES	YES	YES	YES	YES	YES
Standard errors adjusted for school level clustering	YES	YES	YES	YES	YES	YES
General time trend included	NO	YES	YES	YES	YES	YES
Separate trends for low-performing and high-performing students	NO	NO	YES	YES	NO	YES
Separate trends for high poverty and low poverty schools	NO	NO	NO	NO	YES	YES
Student performance-based separate trends for high poverty and low poverty schools	NO	NO	NO	NO	NO	YES
Coefficient on testing	0.028 (0.003)	0.012 (0.003)	0.009 (0.002)	0.019 (0.004)	0.016 (0.002)	0.004 (0.003)
Coefficient on testing x prior year math score	-0.029 (0.006)	-0.039 (0.006)	-0.018 (0.004)	-0.043 (0.008)		-0.012 (0.005)
Coefficient on testing x high poverty school					0.011 (0.003)	0.009 (0.003)
Coefficient on testing x high poverty school x prior year math score						-0.013 (0.005)
Number of counties	2	2	2	2	2	2