Accountability, Ability and Disability: Gaming the System?

David N. Figlio

University of Florida and National Bureau of Economic Research

Lawrence S. Getzler

Virginia Department of Planning and Budget

April 2002

We are grateful to Sheila Murray, Richard Rothstein, and Jim Wyckoff, as well as participants at the APPAM and AEFA meetings for helpful comments and suggestions, and to the National Science Foundation and National Institutes of Child Health and Human Development for research funding. 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.

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; 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.

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. Moreover, we observe that these differential classification responses are particularly pronounced *in grades* in which student test scores are included in the accountability system. 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

[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.

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, and the tests were first used to assess schools and students in 1997-98; hence, for the purposes of this study, 1997-98 is the first relevant year of FCAT testing, and will hereafter be referred to as the first year of FCAT testing.[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. 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 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 our results are due to differences in secular trends in classification of students of different types over time, we further estimate models that isolate the estimated effects of testing on classification *by grade*. This identification strategy should yield even more conservative estimates of the effects of testing on disability classification than the previously-described strategy because it does not capture, say, students reclassified in third grade as a direct result of the testing policy (FCAT testing starts in the fourth grade), or students who remain disability classified following test grades (as declassification is less prevalent than is initial classification.)

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

Our data come directly from the student records of two large herein-unidentified county-level school districts, each among the thirty 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 we have strong evidence to suggest that these districts are still representative of the state of Florida in general for the purposes of this study. We have conducted similar analyses using a more limited dataset and free-lunch classification, rather than student test scores, as a means of identifying potentially low-scoring students, for a set of 23 counties, representing almost two-thirds of the state's population, and found results that were quite similar to those found in identical specifications restricting the sample to the two counties covered in the present study.

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 name, sex, race, and birth date. Students who change school districts over the study period remain in the study provided they relocated to the other district included in the project. We follow from 1995-96 through 1999-2000 every student in these two counties in a district's testing year. (Districts test students in grades not covered by the FCAT.) The set of grades included varies slightly by year, but almost always includes grades three through ten, and in some years begins as early as first grade or ends as late as eleventh grade. School district records include free lunch status, grade, and disability status. In addition, we observe the student's Stanford 9 standardized test score for each student in each observation year. Our sample of 470,747 students includes every student in these two counties in the relevant years. We adjust all standard errors for heteroskedasticity and clustering at the school level.

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

Figure 1 shows the changes in disability classification rates in our population over time. We observe that the rate of disability classification monotonically increased over the time period covered by this study. In each of the two years prior to the first year of effective FCAT testing, 9.1 percent of students in the dataset were classified as disabled. This percentage had increased to 10.4 percent by the 1999-2000 school year.

Figure 2 follows three cohorts of students over time: students in third grade in 1995-96, 1996-97, and 1997-98. The first cohort first faced FCAT testing in fifth grade, the second cohort first faced FCAT testing in fourth grade, and while the third cohort first faced FCAT testing in fourth grade, the testing regime was already in place when they were in third grade. We observe that across cohorts, classification rates increase with grade. However, in the 1995-96 cohort, the major increase in classification rates occurred between fourth and fifth grade (that is, spanning the time following FCAT imposition) while the largest increases in the next two cohorts are between grades three and four. The same patterns are evident in Figure 3, which presents the same data, but only for students scoring in the bottom quarter of the grade three Stanford 9 mathematics test score distribution. In summary, the descriptive evidence supports the notion that schools have classified students, particularly low-scoring students, as disabled more aggressively in the period following the imposition of FCAT testing. Of course, the question remains as to whether these results are causal.

Empirical evidence

Table 1 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. 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 2.6 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 29 percent higher rate of disability classification in the two counties in question.

While schools have an incentive to classify students as disabled regardless of grade, this incentive should be particularly strong for students entering a grade in which student test performance is counted for school accountability purposes. In this case, these grades are the grade levels tested by the FCAT--fourth, fifth, eighth and tenth. Specification 2 from Table 1 investigates whether students in testing grades are differentially likely to be classified as disabled. We observe that, indeed, students in testing grades are substantially more likely to be classified as disabled in the wake of FCAT imposition. Specifically, the effect of test imposition is eight-tenths of a percentage point higher in high-stakes testing grades than in low-stakes testing grades.

Table 2 reports the results of models in which we differentiate the estimated effects of testing for students with different academic or socio-economic backgrounds. Specification 3 is identical to Specification 2 from Table 1, except that all variables are further interacted with free lunch eligibility. We observe that while non-poverty students are no more likely to be reclassified in test grades than in non-test grades, this difference is 0.6 percentage points higher for free lunch-eligible students. This result clearly indicates that low socio-economic-status students are most likely to be reclassified in response to the testing policy.

Specification 4 reports the results of a similar specification, except that all variables are interacted with the student's Stanford 9 mathematics test score from the prior year rather than with free lunch eligibility. The drawback of this exercise is that our first year of analysis is now the 1996-

97 school year, so we 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, particularly in test grades.

The next two specifications of Table 2 repeat the exercise, but constrain the dependent variable to reflect only relatively immutable disabilities. Specifically, in specifications 5 and 6, we estimate the same models as specifications 3 and 4 but where we only count as disabled those certified-disabled students without a learning disability. We observe that the two interaction terms described above (see shaded cells on Table 2) become tiny in magnitude and are not nearly statistically significant. This provides evidence that as we move away from disability classifications that are manipulable, the estimated effects suggestive of manipulation are not present. The final two specifications of Table 2 look only at the most severe disabilities—here, the results are the same.[6] In the case of models where the dependent variable is the presence of a very severe disability—arguably the least manipulable of all—the interaction terms are trivial in magnitude and even more so statistically.

Table 3 reports the results of various models in which students are grouped by school type, rather than test grade. We identify schools as "high poverty" if the school has more than the district-wide median fraction of free lunch-eligible students. We observe (in specifications 9, 10 and 11) that high-poverty schools are significantly more likely to reclassify students than are their relatively low-poverty counterparts. As specification 10 demonstrates, these results are particularly concentrated for free lunch-eligible students. In specification 11, we observe that the three-way interaction between testing, last year's mathematics score, and the high-poverty indicator is negative, is marginally statistically significant.

Specifications 12, 13, and 14 replace the “high poverty” classification with an official state of Florida measure of school performance—the rating of “low performing school” in the 1996-97 school year, the first year of this system. We observe similar patterns. In specification 13, for instance, the effect of increased classification in labeled low-performing schools is actually estimated to be negative, while the three-way interaction with free lunch eligibility is strongly positive, suggesting comparable results to the other set of three-way interactions. Likewise, specification 14 includes a three-way interaction with prior year’s mathematics score. As before, this three-way interaction is negative. Therefore, these results are consistent with a general finding suggesting that high-poverty and low-ability schools tend to differentially classify low-income and poor performing students.

Discussion

We have estimated that the introduction of the high-stakes FCAT testing is associated with a 29 percent higher rate of disability classification.[7] 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 and low-performing 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

[6] We define these disabilities as those that, according to the Florida Education Finance Program, cost the state more than three times what a regular education student would cost. Our empirical findings are not affected by the changes in this “very disabled” definition that we tried.

[7] As mentioned previously, our regressions use student-level data from two large school districts, together representing 26 percent of the state's population.

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[8] 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. 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

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

the minimum percentage still must decrease their percentage of students that are below proficiency by 10 percent when compared with the preceding year.[9]

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

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

[10] 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.

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,” Working paper, University of Florida, 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.

Figure 1: Percentage of students classified as disabled, by academic year

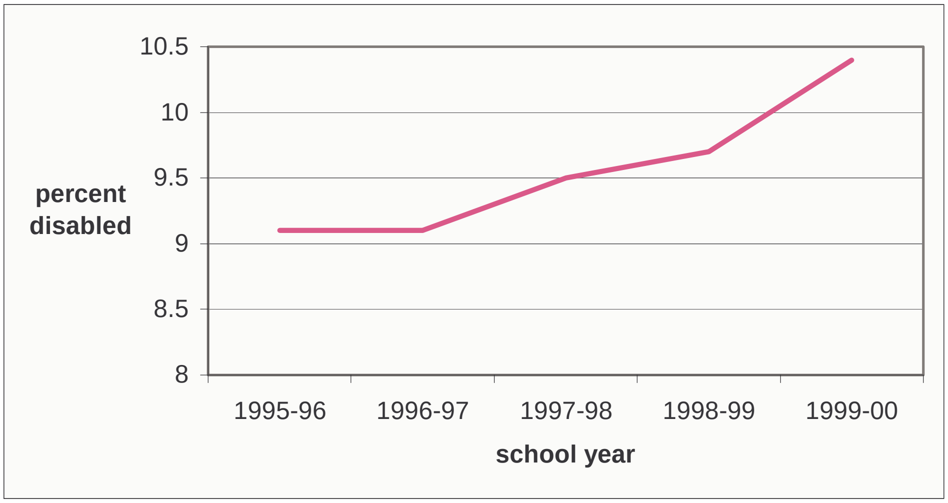

Figure 2: Over-time changes in disability classification, by third-grade cohort

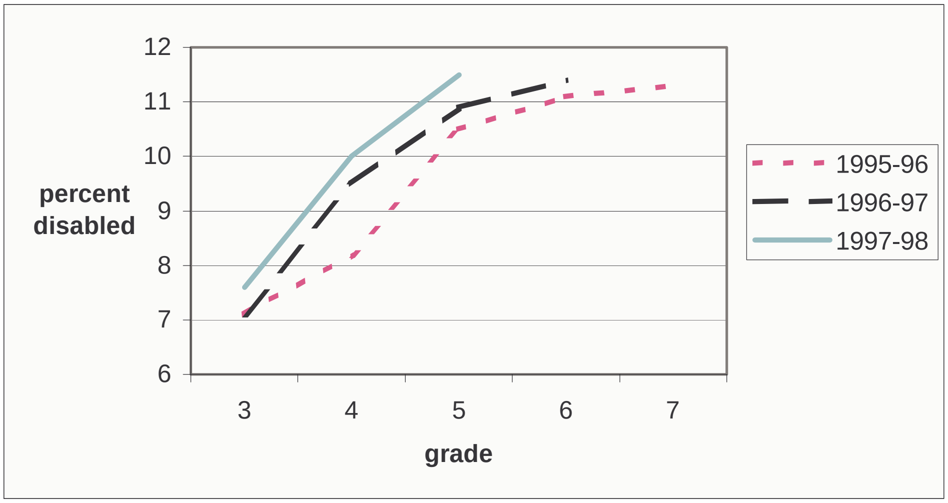

Figure 3: Over-time changes in disability classification, by third-grade cohort; students scoring in the bottom quarter of the grade 3 mathematics test score distribution

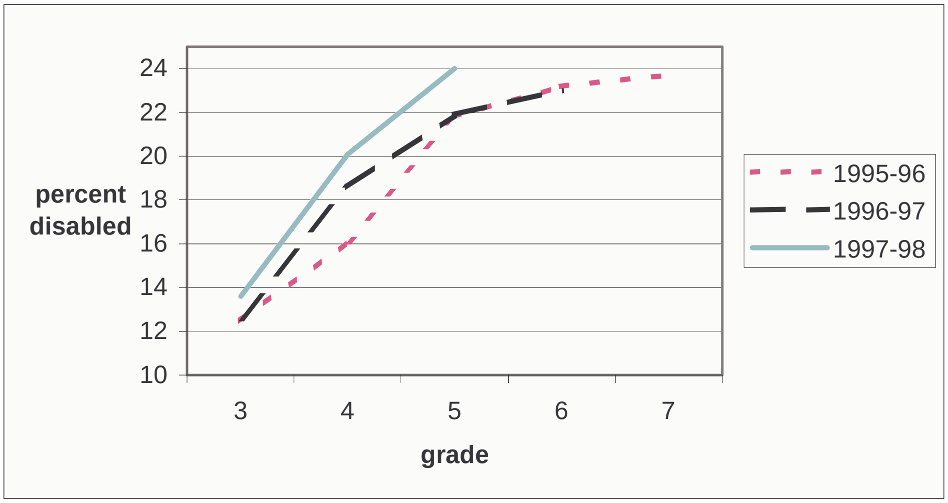

Table 1: Estimated effects of introduction of high-stakes testing on disability classification--Two large Florida counties

Dependent variable: 1=student classified as disabled (in a test-excludable category) by school

Sample: 470,747 students in grades in which students are tested by the district (varies by year, but typically grades 3-10)

Specification	1	2
Student fixed effects?	YES	YES
Grade level effects?	YES	YES
Effect of testing	.026 (.001)	.024 (.001)
Testing x test grade		.008 (.001)

Note: Robust standard errors, reported in parentheses beneath point estimates, are adjusted for error correlations within schools.

Table 2: Differential effects of high-stakes testing on disability classification for students in FCAT grades versus non-FCAT grades

Dependent variable: 1=student classified as disabled (in a test-excludable category) by school

Sample: 470,747 students in grades in which students are tested by the district (varies by year, but typically grades 3-10)

Specification	3	4	5	6	7	8
Student fixed effects?	YES	YES	YES	YES	YES	YES
Grade level effects?	YES	YES	YES	YES	YES	YES
Dependent variable specification	All disabilities	All disabilities	All disabilities except learning disabilities	All disabilities except learning disabilities	Only the most severe disabilities	Only the most severe disabilities
Effect of testing	.007 (.001)	.018 (.001)	.002 (.001)	.002 (.001)	.0001 (.0004)	.0007 (.0002)
Testing x test grade	.001 (.001)	.005 (.002)	-.001 (.000)	-.001 (.001)	-.0003 (.0001)	.0001 (.0003)
Testing x free lunch eligible	.032 (.001)		.009 (.001)		.0012 (.0005)	
Testing x prior year's math test score		.002 (.002)		.002 (.001)		-.0006 (.0004)
Test grade x free lunch eligible	.006 (.001)		.002 (.000)		.0004 (.0002)	
Test grade x prior year's math test score		.006 (.002)		-.001 (.001)		-.0004 (.0003)
Testing x free lunch x test grade	.006 (.002)		.001 (.001)		.0001 (.0003)	
Testing x prior year's math test score x test grade		-.008 (.003)		.001 (.001)		-.0001 (.0005)

Note: Robust standard errors, reported in parentheses beneath point estimates, are adjusted for error correlations within schools. The testing variable represents all years after FCAT was introduced.

Table 3: Differential effects of high-stakes testing on disability classification for students across school type

Dependent variable: 1=student classified as disabled (in a test-excludable category) by school type

Sample: 470,747 students in grades in which students are tested by the district (varies by year, but typically grades 3-10)

Specification	9	10	11	12	13	14
Student fixed effects?	YES	YES	YES	YES	YES	YES
Grade level effects?	YES	YES	YES	YES	YES	YES
Effect of testing	.013 (.001)	.004 (.001)	.009 (.001)	.025 (.001)	.008 (.001)	.016 (.001)
Testing x school is high-poverty	.028 (.003)	.013 (.002)	.017 (.002)			
Testing x school rated low-performing in 1996-97				.005 (.004)	-.009 (.002)	.010 (.002)
Testing x free lunch		.027 (.002)			.035 (.002)	
Testing x school is high-poverty x free lunch		.004 (.002)				
Testing x school rated low-performing in 1996-97 x free lunch					.010 (.004)	
Testing x last year's math score (x100)			.003 (.002)			.003 (.002)
Testing x school is high-poverty x last year's math score (x100)			-.005 (.003)			
Testing x school rated low-performing in 1996-97 x last year’s math score (x100)						-.006 (.003)

Note: Robust standard errors, reported in parentheses beneath point estimates, are adjusted for error correlations within schools.