Education Economics
Vol. 14, No. 2, 235–250, June 2006

Student Assessments, Non-test-takers, and School Accountability

ROBERT J. LEMKE, CLAUS M. HOERANDNER and ROBERT E. MCMAHON

Lake Forest College, USA

ABSTRACT *Much attention has focused recently on using student test scores to evaluate public schools. The No Child Left Behind Act of 2002 requires states to test students and evaluate each school's progress toward having all students meet or exceed state standards. Under the law, however, schools only need to test 95% of their students. When some students do not take the test, variability arises in a school's evaluation as the score of each student who did not take the test remains unknown. Using a statewide assessment administered to 11th graders in Illinois, we investigate this source of variation. In our data, 8% of students do not take the test. By applying a bounding technique to the unknown scores of the non-test-takers, we show that classifying schools as failing or passing against some fixed threshold frequently can be misleading. We also provide evidence that some schools may be strategically selecting some students to not take the test and, by so doing, increasing the school's test scores.*

KEY WORDS: No Child Left Behind; testing; accountability; gaming; education

Introduction

Much attention has focused recently on using student test scores to evaluate public schools. Test scores, it has been argued, can be used by parents, taxpayers, legislators, and educators to hold schools accountable by requiring 'corrective actions' to be taken when 'failing' schools are identified in order to improve public education and make better use of public funds. Corrective actions may include laying-off staff, turning the school over to private or state control, providing vouchers to students, or even shutting down the school. Much hinges, therefore, on the relationship between test scores and the (perceived) quality of schools.

Public Law 107-110, more commonly known as the No Child Left Behind (NCLB) Act of 2002, requires states to test students in the third through eighth grades and once in high school (p. 1450). Although individual test results are intended to help families evaluate their child's progress, the law also evaluates

Correspondence Address: Robert J. Lemke, Box M3, Lake Forest College, 555 N. Sheridan Road, Lake Forest, IL 60045, USA. Tel: +1 847 735 5143; Email: lemke@lakeforest.edu

0964-5292 print/1469-5782 online/06/020235-16 © 2006 Taylor & Francis
DOI: 10.1080/09645290600622970

schools. NCLB requires districts to produce annual local report cards for each school, specifying average student performance on state assessments. Under NCLB, by the 2013–14 academic year 'all students … will meet or exceed the state's proficient level of academic achievement' (pp. 1447–1448). Moreover, not only must schools demonstrate substantial improvement each year, but progress must be demonstrated separately for 'economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency' (pp. 1446–1447).

As is common with federal mandates aimed at accountability, there are some caveats.[1] Not only do states set their own standards (subject to a federal approval process),[2] but schools can also administer different tests to different groups of students (e.g., to the disabled). Moreover, NCLB does not, in fact, require schools to test all of their students: '… for a school to make adequate yearly progress, … not less than 95 percent of each group of students … who are enrolled in the school are required to take the assessments' (p. 1448).

Under NCLB, whether a school is deemed to be making adequate yearly progress depends on the percentage of its students who meet or exceed state standards of proficiency. When some students do not take the test, however, variability arises in each school's evaluation as the scores of those students remain unknown.[3] As NCLB ties substantial rewards and severe punishments to test outcomes, this variability in test scores may lead to an unequal treatment of schools. Moreover, because of the overwhelming importance of test scores, teachers and school administrators may be tempted to 'game the system' in order to increase the scores of their students. There are many ways to game this system. Students could be held back a grade to postpone testing. Rules could be relaxed to encourage the weakest students to drop out of school. Borderline students could be reassigned to be tested with the learning disabled or with those students classified as having limited English proficiency. A well-rounded curriculum could be substituted for a narrow curriculum that teaches to the test. Or teachers and administrators may simply provide the answers to the students.[4]

The purpose of this paper is to determine the ability of test scores to accurately classify and compare schools when not all students are tested with the primary assessment. We also investigate whether schools appear to be gaming the system by allowing their worst students to not be tested or to be tested using a modified assessment.

We use school-level data from May 2002 in Illinois. A particularly attractive feature of the data is that each school reports in May the number of its students who are subject to state testing and the number who were actually tested. This allows for an accurate measure of the number of students who were not tested. Comparatively, other data sets typically provide fall enrollments and spring test numbers.

In May 2002, 8% of 11th graders in Illinois did not take the state's primary high school assessment. By applying a bounding technique to the unknown scores of the non-test-takers, we show that classifying schools as 'failing' or 'passing' against some fixed threshold can frequently be misleading. For example, if the standard for a school to pass is that at least 60% of its students meet or exceed Illinois standards, then 194 of our 573 schools would pass if non-test-takers are ignored, as is standard procedure under NCLB. On the other hand, 36 of these 194 schools—almost one in five—would have failed if their non-test-takers had all taken and failed the test, while 41 failing schools would have passed the standard

if all of their non-test-takers had taken and passed the test. We also use school level regressions to demonstrate that schools may be selecting some of their students to not take the state's primary assessment in order to increase their pass rates.

In the next section we discuss some of the institutional features of NCLB and testing in Illinois high schools. The data are introduced in the third section, while the results concerning the classification of schools as passing versus failing depending on the treatment of those students who do not take the primary assessment are presented in the following section. Regression analysis is used in the fifth section to provide evidence that schools might be gaming the system in order to increase test scores. A discussion of the results and policy implications is offered in the final section.

NCLB and the Prairie State Achievement Examination

NCLB was signed into law by President Bush on 8 January 2002. This action was a continued response to the 1983 *A Nation at Risk* report, which documented large gaps in the academic achievement levels of students across socio-economic and racial/ethnic backgrounds. A cornerstone of NCLB requires all students by the 2013–14 academic year to meet or exceed state proficiency standards approved by the US Department of Education. In theory, once all students meet or exceed the standards, the gaps in minimum acceptable achievement standards across different groups of students will be eliminated. Under NCLB, all states were also required to submit a proposal to the US Department of Education specifying their plans for testing students and documenting their standards, annual improvement goals, and plans for spending federal funds on improving overall student achievement. Similar to 1996 welfare reform, NCLB increased funding to states in exchange for accountability. In order for a school to demonstrate adequate yearly progress, it must demonstrate overall improvement as well as yearly improvement separately by student race, ethnicity, disability, English ability, and economic status. Each state plan must also include a series of corrective actions to be undertaken when a school or district fails to meet its annual improvement goals.

At the time NCLB became law, many of its goals and requirements were already present in Illinois. The state of Illinois has been measuring student progress with standardized tests since the early 1980s, and continues to do so in Grades 3–8 with the Illinois Standards Achievement Test (ISAT) and in Grade 11 with the Prairie State Achievement Examination (PSAE). Also in accordance with NCLB, each public school district in Illinois uses its ISAT and PSAE scores to publish annual school report cards. Illinois also has in place a detailed six-year plan for improving under-performing schools, which, among other things, includes provisions for 'deferring programmatic funds or reducing administrative funds, implementing a new curriculum (with professional development), replacing personnel, establishing alternative governance arrangements, appointing a receiver or trustee to administer the district in place of the superintendent and school board, or abolishing or restructuring the school district. The state may also authorize students to transfer to higher performing public schools operated by another school district … [or create] a charter school' (Illinois State Board of Education [ISBE], pp. 6–7, 2002).

Illinois makes special provisions for students with disabilities and for students with limited English proficiency. Each disabled student is enrolled in

an Individualized Education Program (IEP) and taught according to Illinois learning standards but possibly without a grade level, depending on the severity of the disability. Because such a student is not associated with a grade level, the student's IEP team (education experts, teachers, administrators, and parents) decides at what level the student will be tested and whether the student will take the standard assessment (ISAT or PSAE) or a modified assessment, namely the Illinois Alternate Assessment (IAA). Similarly, the language and academic skills of students with limited English proficiency are measured not with the ISAT or the PSAE, but with another modified assessment called the Illinois Measure of Annual Growth in English Examination (IMAGE).[5]

The Data

Every school district in Illinois must provide the ISBE with its underlying data from which annual local area report cards are generated. Some variation of these data is then made available to the public. In addition to student demographics and school and staffing characteristics, the 2001–02 school-level data reports May 2002 test scores along with some information about the number of students who were tested.[6] It is the only year for which the ISBE reports the necessary detailed information on the number of test-takers. As we are interested in studying the effect that not taking the primary state assessment has on the evaluation of schools, we must know the number of students who are designated by the state to take the test and the actual number of test-takers. A particularly attractive feature of these data is that schools are asked in May to report both enrollment and test data. This is far superior to using a data set that contains fall enrollments and May tests as it avoids needing to make heroic assumptions concerning student mobility during the school year. Specifically, every school reports its May 2002 Grade 11 enrollment, PSAE scores, the percentage of students who were not tested with any state assessment (including the IAA and IMAGE), and the total number of PSAE tests that were scored.[7] Thus, each school's May Grade 11 enrollment reflects the number of students designated by the state to take the PSAE, and the number of PSAE tests scored reflects the actual number of students tested. Unfortunately, the number of scored PSAE tests is not reported for other years. This prevents us from carrying out time series analysis or including school fixed-effects.

As the PSAE is given only to 11th graders, our unit of analysis is a high school. Table 1 provides the descriptive statistics for our final set of 573 high schools.[8] The mean and standard deviation of each variable are reported twice, first when the school is the unit of observation and again when the data are weighted by the number of students. For example, the average school had about 52.5% of its tested 11th graders pass the PSAE, whereas about 53.6% of all 11th graders who took the PSAE in May 2002 actually passed it.[9] Pass rates from May 2001 were about the same.

The next four rows of Table 1 report descriptive statistics concerning test-takers versus non-test-takers. Almost 92% of all 11th graders took the PSAE, leaving about 8% of such students who did not take the state's primary assessment. Of the roughly 8% of students who failed to take the PSAE, some avoided all state assessments (5.9 of the 8 percentage points) while the remainder were tested with an alternative state assessment (2.1 of the 8 percentage points).

A brief comment on how the percentage of test-takers was calculated is warranted. Each school reported how many students were subject to standardized

Table 1. Descriptive statistics: Illinois High Schools 2001–02 school year

	Weighted by school		Weighted by student			
	Mean	Standard deviation	Mean	Standard deviation	Minimum	Maximum
Percentage who passed the PSAE in May 2002	52.513	16.798	53.557	19.803	4.3	90
Percentage who passed the PSAE in May 2001	52.440	17.083	53.252	19.823	3.1	89.5
Percentage of 11th graders who took the PSAE in May 2002	94.164	7.542	91.957	8.171	45.0	100
Percentage of 11th graders who did not take the PSAE in May 2002	5.836	7.542	8.043	8.171	0	55.0
Percentage who were not tested in May 2002	4.493	6.592	5.913	6.890	0	46.8
Percentage who took an alternative assessment in May 2002	1.343	2.562	2.131	3.284	0	24.1
Percentage of students who are white	79.030	31.019	66.649	32.761	0	100
Percentage of students who are black	12.039	25.511	16.582	27.078	0	100
Percentage of students who are Hispanic	6.922	15.046	12.815	20.150	0	97.5
Percentage of students from low income households	24.592	24.135	26.327	27.594	0.3	100
School enrollment	916.0	837.5	1 680.4	909.3	19	4 235
Annual mobility rate	30.055	24.027	30.559	28.126	0.7	96.5
Chronic truancy rate	3.497	5.881	3.572	5.820	0	62.3
Inner-city school	0.208	0.406	0.322	0.468	0	1
Suburban school	0.611	0.488	0.623	0.485	0	1
Rural school	0.182	0.386	0.056	0.230	0	1
Years of experience for the average teacher (district)	15.075	2.175	14.759	2.204	6.6	21.6
Percentage of teachers with at least a Masters degree (district)	40.939	18.061	52.419	16.525	4.9	82.9
Number of high schools in the district	6.913	16.686	9.940	19.606	1	57
School is in a High School district	0.244	0.430	0.419	0.494	0	1
Average teacher salary (district)	47 388	11 717	55 212	12 239	28 689	87 677
Average administrator salary (district)	82 541	15 970	92 463	16 307	46 575	133 609

Source: Illinois State Board of Education. Number of observations = 573. With reference to standardized tests, a student is defined as passing the test if he/she meets or exceeds the state standards for the test. Schools not in a high school district are in unit districts that encompass K–12 grade levels.

testing in May 2002, the percentage of such students who failed to be tested, and the number of students who took the PSAE.[10] Using the percentage of students not tested and the total number who were subject to state testing allows for the number of students who were not tested with any assessment to be calculated. The number of students who were tested by an alternative assessment is then simply the

difference between the total number of students subject to testing and the number of students who took the PSAE or did not take any test. Of the 573 schools, 153 report all of their students took the PSAE. In another 164 schools, the percentage of students who were not tested by any assessment explained the entire gap between the number of students subject to testing and the number of PSAE test-takers. Thus, 153 schools (26.7% of the sample) are associated with 100% of its 11th graders taking the PSAE, and 317 schools (55.3% of the sample) are associated with no students taking an alternative assessment.

The remaining rows of Table 1 present school characteristics. The variables listed in the last six rows of Table 1 are reported by the ISBE at the district level rather than the school level.[11]

The Effect of Non-test-takers on School Pass Rates

As mentioned in the previous section, the number of 11th graders who did not take the PSAE in May 2002 is not trivial. On average, 80 out of every 1000 11th graders did not take the test. This statistic, however, masks a great disparity across schools. All 11th graders took the PSAE in 153 of our 573 high schools, while over 9% of 11th graders did not take the test in the remaining 420 schools.[12]

In order to examine whether students who do not take the primary assessment (those we call non-test-takers) have a substantial effect on the classification of schools as passing or failing, we undertake a bounds analysis.[13] We start by considering the number of schools that would be reclassified under the most extreme assumptions possible—namely, assuming that all non-test-takers would have failed had they taken the test versus assuming that all non-test-takers would have passed had they taken the test. Table 2 presents these results.[14]

The first column of Table 2 gives the classification requirement so that in the third row a school passes the state standard if at least 60% of its students pass the PSAE. The third column in Table 2 reports the number of schools that pass when, as is standard practice, non-test-takers are ignored from the calculation of a school's pass rate. In row three, for example, only 194 of the 573 schools (33.9%) would pass a state bar set at 60% assuming there is no self-selection among the non-test-takers.

Table 2. Passing versus failing schools: accounting for the non-test-takers

School passes if at least this percentage of students passes the PSAE	Number of schools that would pass if all non-test-takers are assumed to *not pass* the PSAE	Number of passing schools when non-test-takers are ignored	Number of schools that would pass if all non-test-takers are assumed to *pass* the PSAE	Fraction of all schools for which assumptions on non-test-takers matter
40%	467	479	505	0.0663 (0.0104)
50%	349	385	418	0.1204 (0.0136)
60%	158	194	235	0.1344 (0.0142)
70%	34	58	75	0.0716 (0.0108)

Note: The May 2002 PSAE test scores from 573 high schools are considered. Of these, 153 (26.7%) report that all of their 11th graders took the test, and therefore cannot be reclassified as passing or failing depending on the treatment of non-test-takers. The fraction of schools for which the assumptions matter is statistically significant at all conventional significance levels. (The appropriate standard error is reported in parentheses in the fifth column.)

The second and fourth columns of Table 2, in contrast, show the worst and best classification of schools given the unknown scores of the non-test-takers. Assume, for example, that everyone who did not take the PSAE would have failed had they taken it. These hypothetical failing scores can then be used to recalculate each school's pass rate, which necessarily (weakly) reduces the number of passing schools. At the other extreme, assume everyone who did not take the PSAE would have passed had they taken it. These hypothetical passing scores can then be used to recalculate each school's pass rate, which necessarily (weakly) increases the number of passing schools.[15]

When the state bar for passing is set at 60%, 194 schools pass by having at least 60% of their students who take the PSAE pass it. If, however, non-test-takers are assumed to fail, then only 158 schools pass. Thus, the worst possible assumption reclassifies almost one in every five previously passing schools as now failing. At the other extreme, 235 schools pass if non-test-takers are assumed to pass. In total, 77 schools, or 13.4% of the entire sample, are affected by the assumptions of the abilities of the non-test-takers. Recall, however, that 153 schools reported 100% testing and therefore cannot be affected (in either direction) by the assumptions concerning non-test-takers. In this light, the 77 affected schools represent over 18% of the possibly affected 420 schools. Similarly, 69 schools are affected when the bar for passing is set at 50%. In this case, 385 schools pass when non-test-takers are ignored, but only 351 pass if non-test-takers are assumed to fail while 418 schools pass if non-test-takers are assumed to pass. Although the percentage of schools affected by the treatment of non-test-takers is greatest when the bar is set at 50–60%, many schools are also affected at a bar of 40% or 70%. If, as required by NCLB, the classification of schools by test scores is used for policy prescriptions such as funding, invoking a voucher system, laying-off staff, and so on, the existence and potentially unequal treatment of non-test-takers might lead to an unequal treatment of schools based not on successful teaching (or even testing ability) but on possibly manipulable student behavior towards test taking.

According to Table 2, the classification of about 7–13% of the schools is sensitive to the assumptions on non-test-takers when the bar for passing is set anywhere between 40% and 70%. It is important to note, however, that this is not the result of 7–13% of the population of schools having extremely high rates of not taking the test (and therefore being sensitive to the assumptions at all thresholds), but rather it comes about because many schools are sensitive to the assumptions at some threshold. In fact, almost 82% of the schools have fewer than 10% of their 11th graders not take the PSAE.

It is also of interest to know how susceptible school classifications would be to administrators gaming the system by encouraging up to 5% of its worst students to not take the test. Suppose every school that does not currently have at least 5% of its 11th graders not take the PSAE starts to encourage 5% of its worst students to not take the test. This will falsely increase the school's reported pass rate. Under this hypothetical change, and assuming the bar is set at 60%, 228 schools pass as compared with 194 passing schools in Table 2. Thus, 34 schools could increase their pass rate from somewhere less than 60% to somewhere more than 60% simply by having a small portion of its 'bad' students not take the PSAE. In this case, almost 22% of schools would be sensitive to the treatment of non-test-takers, which is almost 65% more than the 13.4% of schools that were sensitive to the treatment of non-test-takers at the same bar in Table 2.

Gaming the System

In the previous section, we showed that the classification of schools as passing versus failing is sensitive to the treatment of those students who do not take the primary assessment. In this section, we further investigate whether schools might be gaming the system by having their weakest students not take the PSAE. There are two ways in which our data could reveal gaming. First, we would expect schools with the most to gain to be the most willing to game the system. Therefore, we investigate whether schools with lower pass rates in May 2001 have a greater fraction of its 11th graders not take the PSAE in the following year. Second, if a school is successful in having some of its worst students not take the PSAE, this should reveal itself in a higher pass rate. Thus, we investigate whether schools with higher rates of not taking the test achieve higher pass rates than comparable schools. Moreover, we are also in a position to further dissect the means by which schools might be gaming—namely, by not having some students tested versus reclassifying some students so they take an alternative assessment.

Although we find support for both relationships, and in particular the relationship is most evident by schools simply not testing some students rather than by reclassifying students to take the IAA or IMAGE, it should be noted that schools might actively game the system in such a way that does not involve reassigning test takers in this way. Chicago Public Schools, for example, have relaxed the requirements for dropping out of high school. By reducing the barriers to dropping out for students aged 16 or older, Chicago will probably experience more drop-outs, which leaves them with fewer 11th graders they are required to test. Assuming that high school drop-out are less likely to pass the PSAE than non-drop-outs and would have been more likely to not take the PSAE had they been enrolled, such a policy will have two effects. Chicago schools will have higher rates of test taking *and* higher than expected pass rates. Thus, finding evidence of gaming the system using rates of test taking is, *a priori*, a difficult task.

In a regression of the percentage of students not taking the PSAE on the previous year's pass rate, evidence of gaming will be revealed in a negative relationship. That is, schools with lower pass rates last year are assumed to have a greater incentive to game the system and, therefore, to test fewer of its students currently.[16]

In column (1) of Table 3, we report ordinary least squares (OLS) estimates and standard errors from regressing the percent of 11th graders not taking the PSAE in May 2002 on the school's pass rate from the previous year (and its square) plus a variety of school characteristics.[17] We include the square of previous test scores to allow for a non-linear relationship. In particular, the schools with the lowest pass rates are the ones that potentially benefit the most by gaming. Thus, the marginal effect of increased pass rates on rates of non-test-taking will probably be different across schools.[18] The model also includes the percentage of students who are black and Hispanic to capture racial and ethnic effects, school enrollment (and its square) to capture size effects, the percentage of students from low-income households, the mobility rate, and the truancy rate to capture household differences and student attachment to the school, and dummy variables indicating if the school is in a primary city or a suburban county of a metropolitan statistical area (MSA) to capture location effects.

In model 1, the dependent variable is the percent of 11th graders who do not take the PSAE for any reason. The coefficient on the linear term of May 2001 pass

Table 3. OLS regressions: predicting the percentage of test-takers from the previous year's test results

Percentage of 11th graders in May 2002 who:	did not take the PSAE [Column (1)]	did not take any test [Column (2)]	took an alternative assessment to the PSAE [Column (3)]
PSAE pass rate in May 2001	−0.3733*	−0.3553*	−0.0180
	0.0774	0.0694	0.0325
PSAE pass rate in May 2001 squared	0.0030*	0.0030*	−0.00002
	0.0006	0.0006	0.0003
Percentage of students black	−0.0228	0.0003	−0.0225*
	0.0187	0.0168	0.0078
Percentage of students Hispanic	0.0839*	−0.0086	0.0925*
	0.0229	0.0206	0.0096
School enrollment	0.0045*	0.0030*	0.0015*
	0.0010	0.0009	0.0004
School enrollment squared	−0.0010*	−0.0006*	−0.0003*
	0.0002	0.0002	0.0001
Percentage of students from low-income households	0.0943*	0.0321	0.0654*
	0.0364	0.0326	0.0153
2001-02 Mobility rate	−0.0128	0.0410	−0.0538*
	0.0300	0.0265	0.0124
2001-02 Chronic truancy rate	0.1848*	0.2276*	−0.0428*
	0.0485	0.0434	0.0203
School is in an inner-city of an MSA	0.2274	0.8537	−0.6264
	1.1893	1.0656	0.4992
School is in a suburban county of an MSA	0.2525	0.4784	−0.2260
	1.1028	0.9881	0.4629
Constant	10.6912	9.3095	1.3817
	2.8965	2.5953	1.2157
R-squared	0.5432	0.4801	0.5009
Number of observations	573	573	573

Note: Standard errors are reported beneath the coefficient estimates. All regressions are weighted by the number of test-takers in each school to correct for heteroskedasticity.
*Statistically significant at the 1% level.

rates is strongly negative, but the coefficient on the squared term suggests that predicted pass rates do eventually increase with May 2001 pass rates. In particular, the relation is negative for schools with May 2001 pass rates below 62%.

The remaining two models in Table 3 repeat the analysis by altering the dependent variable. In column (2), the dependent variable is the percentage of 11th graders who do not take any state assessment (i.e., they do not take the PSAE, IAA, or IMAGE). Again, a very strong negative relationship is found between previous pass rates and current rates of not being tested. In this case, the negative relationship persists for schools with May 2001 pass rates of 59% or less. In column (3), the dependent variable is the percent of 11th graders who took an alternative assessment (i.e., the IAA or IMAGE) in place of the PSAE. In this case, there is no statistical relationship between previous pass rates and current rates of taking an alternative assessment.[19] Consistent with these results, therefore, is that the worst performing schools simply tend to test fewer of their students rather

than actively reassigning students as being disabled or having limited English proficiency.

These results are best interpreted as measuring steady-state relationships. An alternative specification would be to consider the effect *changes* in pass rates have on the *change* in the percent of students not taking the test. The 2000–01 data available from the ISBE, however, do not report the same information on the percentage of test takers and the number of students who took the PSAE. At best a rough calculation can be made using sparse data on students taking the IAA or IMAGE. Estimating the same models, except using the change in pass rates and change in rates of not taking the PSAE, reveal the same general patterns but with weaker statistical significance.

We now turn to the question of whether schools with more students not taking the test experience pass rates higher than expected. We investigate this relationship by regressing pass rates on the fraction of students not taking the test, whereby a positive coefficient on the percentage of students not taking the test would be consistent with gaming.

There are two major differences between the framework for this question and the previous one. First, as we are now predicting pass rates, we want to include teacher inputs into the production function. We do this by including average teacher experience and the percent of teachers with a Masters degree, both measured at the district level. Second, as Hoxby (2000) noted, using school-level data to identify cross-district output effects can produce biased estimates as parents do not randomly distribute themselves across school districts, and thus school inputs are not likely to be exogenous to the education process. We address this endogeneity problem with instrumental variables. In particular, we are concerned with the three variables most in the school's control—rates of not taking the test, teacher experience, and teacher education. We instrument for these three variables with the number of high schools in each school district, whether the school is in a high school district (9–12) versus being in a unit district (K–12), average teacher salary, and average administrator salary.

The endogeneity problem arises because parents may make housing decisions based on school inputs. If so, then school inputs are not exogenous to the education process. Even in this case, however, location decisions should be made taking into account school, not district, inputs/quality. The number of high schools in a district, therefore, should be of little importance to parents compared with the actual inputs that different schools offer. Likewise, whether a particular school is administered as part of a high school district or as a unit district should be of little importance to the location decision as the per-student funding process is the same. Consequently, the number of high schools and district type should be uncorrelated with location decisions, and thus lend themselves as instruments.

We use average teacher salary and average administrator salary as instruments as well, although for a different reason. There is a large literature concerning whether higher teacher salaries actually purchase higher quality teachers.[20] If money matters in terms of which resources additional finances can purchase (such as more experienced teachers) but does not matter directly in terms of salaries, then average teacher and administrator salaries are valid instruments for education inputs.

The regression results using these instruments are presented in Table 4.[21] In column (1) of Table 4, the percentage of students who passed the PSAE in May 2002 is regressed on the percentage of 11th graders who did not take the PSAE in

Table 4. IV regressions: predicting pass rates on the May 2002 PSAE from the percentage of test-takers

	Column (1)	Column (2)	Column (3)
Percentage of 11th graders who did not take the PSAE in May 2002	2.8192** 1.2290		
Percentage of 11th graders who did not take any test in May 2002		1.6085* 0.5206	
Percentage of 11th graders who took an alternative assessment in May 2002			–2.3083** 0.9971
Average years experience by teachers (district)	0.6626	–0.0146	–1.0431**
	1.2474	0.7289	0.5151
Percentage of teachers with at least a Masters degree (district)	–0.0360	0.2117**	0.4800*
	0.2289	0.1064	0.0867
Percentage of students black	–0.5670*	–0.3306*	–0.1617
	0.1114	0.0581	0.1105
Percentage of students Hispanic	–0.2687*	–0.3118*	–0.3652*
	0.0645	0.0396	0.0379
School enrollment	–0.0085**	–0.0071*	–0.0055**
	0.0042	0.0027	0.0022
School enrollment squared	0.0033*	0.0026*	0.0019*
	0.0010	0.0006	0.0005
Percentage of students from low-income households	–0.7996*	–0.5303*	–0.2514*
	0.2116	0.0848	0.0793
2001–02 mobility rate	0.1358	0.1140	0.1352†
	0.1262	0.0862	0.0748
2001–02 chronic truancy rate	–0.9251*	–0.6849*	–0.3322*
	0.3238	0.1681	0.0900
School is in an inner-city of an MSA	5.8762	3.5815	1.3479
	4.3863	2.7462	2.3415
School is in a suburban county of an MSA	1.5507	0.9350	0.3470
	3.5331	2.3096	1.9281
Constant	55.3680	56.7030	61.9272
	12.5004	8.2195	6.6425
Baseman statistic (*p* value)	0.8819	0.1927	0.0013
R-squared	0.1609	0.6379	0.7492
Number of observations	573	573	573

Note: The dependent variable is the percentage of students who passed the test in May 2002. Standard errors are reported beneath the coefficient estimates. All regressions are weighted by the number of students who took the PSAE to correct for heteroskedasticity. For each regression, the percentage of students taking (or not taking a test), average teacher experience, and percentage of teachers with at least a Masters degree are instrumented with the number of high schools in the district, whether the school is in a High School district, average teacher salary, and average administrator salary.
*Statistically significant at the 1% level.
**Statistically significant at the 5% level.
†Statistically significant at the 10% level.

May 2002.[22] A statistically positive relationship between test-takers and test results is found, suggesting that schools with a smaller fraction of students taking the test are associated with higher pass rates. In terms of the point estimate, a school's pass rate is expected to increase by over 2 percentage points for every additional 1% of students who do not take the PSAE. It should be noted that not only is the direction of the impact consistent with gaming, but moreover the magnitude of the effect is statistically no different from unity. That is, if every student who does not take the PSAE would have failed it, then the coefficient on the percentage of students not taking the PSAE should be one, and this cannot be rejected at any standard significance level.

In column (2) of Table 4, pass rates are regressed on the percentage of students who are not tested at all. Again, a statistically positive relationship that is not statistically different from one is found. Only in column (3), when the explanatory variable is the percentage of students who take an alternative assessment, is the relationship negative.[23] This pattern is consistent with some schools strategically choosing certain students to not be tested, possibly by encouraging absences on the test day or by not requiring their worst students who were absent on the test day to take the make-up test. However, there is no evidence that schools successfully increase their pass rates by reassigning students to take alternative assessments.

Discussion

Students in public schools have been subject to standardized testing for decades. With the passage of NCLB, however, teachers, schools, and administrators will also be held accountable for the results of those tests. Yet, the ability of standardized tests to accurately reflect school performance remains in doubt (Haney, 1993, 2000; Hillocks, 2002). Some argue standardized tests have a cultural bias, while others question the accuracy of such tests when students have very little, if any, incentive to do well on the tests.

The most egregious problems with standardized testing, however, may concern the incentives faced by school staff and administrators. In order to keep test scores high, there is an incentive for schools to classify students as learning disabled if doing so allows the student to take a modified (easier) test or not be tested at all (Figlio and Getzler, 2002; Jacob, 2002). Districts may relax their dropout procedures. Teachers may find themselves teaching to the test in place of developing deeper analytical skills or creativity (Hillocks, 2002; Kane and Staiger, 2002). Students who are thought to be unlikely to score well on the test and who were absent on the testing day may not be actively pursued by the school to take the test during the make-up period. If only certain grade levels are tested, schools may choose to hold back students to postpone their being tested for a year (Haney, 2000; Lewin and Medina, 2000).[24] And there have been allegations that teachers game the system by providing questions or answers to students before the test or by altering student test forms (Archibold, 1999; Goodnough, 1999; Jacob, 2002; Jacob and Levitt, 2003a, 2003b; Wilgoren, 2001).

The central focus of this paper concerns one way in which schools can game the system and will be able to continue to do so under NCLB—namely, by having a substantial portion of students not take the primary test. In Illinois, 8% of all 11th graders failed to take the PSAE, Illinois' primary assessment for 11th graders, in May 2002. Some of these students took other tests designed for students with

disabilities or limited English proficiency, while others simply were not tested. Under NCLB, schools can continue to administer different tests to different groups of students and test only 95% of their students.

We have shown that when a substantial number of students fail to take the test, accurate comparisons across schools cannot be made. If schools are classified as passing if at least 60% of their students meet or exceed state standards, for example, then over 13% of the schools in our sample could be classified as both passing and failing, depending on what one assumes about the abilities of the students who did not take the test. Furthermore, if schools would decide to take full advantage of the 95% rule under NCLB, almost 22% of all schools could be mistakenly labeled as passing state standards. This is particularly troublesome as NCLB ties substantial rewards and harsh punishments to test scores.

Finally, regression results are consistent with schools strategically choosing some students to not take the test in order to increase their pass rates. First, schools with low pass rates allow more of their students to not take the PSAE the following year. Second, there is strong statistical evidence that having fewer students take the PSAE is associated with higher pass rates. Both of these results hold when considering students who did not take any state assessment; however, there is little evidence that either relationship is fostered by schools purposely reassigning marginal students as being disabled or as having limited English proficiency in order for these students to qualify for a modified assessment.

These results highlight several policy implications if public sentiment requires using annual average test scores to evaluate schools. One possibility is to require all students to take an identical test. Schools can then be held up to the same standard as they are all measured against the same test. The evaluation of individual students could, of course, take into account disabilities or English proficiency. Although we find little evidence that schools gamed the system in May 2002 by reclassifying students in this way, the possibility exists and will probably be more enticing when the 95% testing rule of NCLB is enforced.

A more feasible policy change, however, would be to make it standard practice to assume that non-test-takers would have failed the test. Under this practice, schools would always have an incentive to test as many students as possible. Legislation such as NCLB could then require schools to attain a 95% pass rate instead of imposing a 95% testing requirement. This would continue to allow schools to not test some of their students for legitimate reasons, but removes negative incentives to have their worst students not take the test. Even this policy, however, is manipulable as schools may try to adjust who is test eligible by relaxing drop-out rates, increasing expulsions, or holding students back a grade.

Acknowledgements

Financial assistance has been provided by the Richter Apprentice Scholars Program at Lake Forest College. The authors thank Louis Cain, Kelly DeRango, Stephen Drinkwater, Richard Dye, David Harrington, Stephen Karlson, John Pepper, Lisa Powell, John Karl Scholz, Jeff Sundberg, Rob Witt, and seminar participants at Loyola University Chicago, Surrey University, and the 2004 Midwest Economics Association Meetings for valuable comments. Any remaining errors are the authors' own.

Notes

1. For example, the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 mandated time limits and work requirements on recipients of temporary aid for needy families. The legislation, however, also allowed each state to exempt up to 20% of its average monthly enrollment for personal hardship considerations (Public Law 104-193 , p. 2138).
2. Recently, Texas lowered its standards in order to meet the NCLB requirements (Dillon, 2003).
3. We are not in a position to discuss the ability/accuracy of standardized tests to measure student achievement or to facilitate comparisons across schools or districts. Rather, we take as given the accuracy of the test and, instead, focus exclusively on the statistical effect non-test-takers have when comparing average test scores across schools. For a broader discussion of accountability systems and their potentially perverse incentive structures, see Kane and Staiger (2002). For a discussion of the effectiveness of standardized testing, see Dylan (2001).
4. Each of these forms of gaming has been exposed in various schools. For some particular examples and a more in-depth discussion of gaming the system under high-stakes testing, see Archibold (1999), Figlio and Getzler (2002), Goodnough (1999), Jacob (2002), Jacob and Levitt (2003a, 2003b), and Wilgoren (2001).
5. A high school can satisfy the NCLB requirement of testing at least 95% of its students by having students take either modified assessment. Thus, even under NCLB, it will be common to have fewer than 95% of 11th graders in a high school take the standard assessment (i.e., the PSAE in Illinois).
6. Links to the data files can be found online (http://www.isbe.net/research/reports.htm#Statistics).
7. Although test scores are also reported for the ISAT, the number of ISAT test-takers is not reported. Thus, we cannot execute a similar analysis for the ISAT. We also do not have IAA or IMAGE scores for all schools as schools only report these scores if more than 5% of its students take either assessment.
8. Twenty high schools are dropped for not reporting PSAE scores for 2000–01 or 2001–02, and the 48 high schools that report ISAT scores are also omitted from the analysis. Most of these 48 schools are combination Junior/Senior High schools. Nine schools were also dropped for having incomplete school characteristic data or having more than 250% annual growth in the number of eligible test takers from 2000–01 to 2001–02.
9. A student passes the PSAE if he/she meets or exceeds the state standards. Likewise, the percentage of students in a school that meet or exceed state standards is said to be the school's pass rate.
10. Schools actually report the percentage of students who were not tested in reading and the percentage who were not tested in mathematics. In most cases these percentages were identical or nearly identical. They differed by at most 8.7 percentage points. We always took the smaller of the two percentages. In this regard, our procedure offers a conservative estimate on the number of students who were not tested at all and, by default, liberally classifies some students as having taken an alternative assessment.
11. The mobility rate is the percentage of students who moved into or out of the school at any time during the school year. With few exceptions, Illinois arranges school districts into three types—unit (K–12), elementary (K–8), and high school (9–12). Just less than 25% of our schools are in high school districts.
12. Having all 11th graders take the PSAE is somewhat suspicious given that modified tests are available and that the daily attendance rate is below 100% for every public high school. Schools, however, are allowed and encouraged to administer make-up tests. In any case, schools that report having tested all of their 11th graders will bias the data in favor of accepting the results from standardized testing at face value.
13. Manski (1995) provides a thorough treatment of this kind of bounds analysis.
14. There is a recent literature on policy evaluation when this type of (unobserved) counterfactual problem exists. See, for example, Heckman *et al.* (2002), Manski *et al.* (2002), and Pepper (2003).
15. For example, consider a school with 200 students, 40 of whom did not take the PSAE. Of the 160 students who took the test, 104 students passed. Under the assumption of no self-selection, this school's pass rate is 65%. If all 40 non-test-takers are assumed to fail the test, the school's pass rate is 52% (104 out of 200). If all 40 non-test-takers are assumed to pass the test, the school's pass rate is 72% (144 out of 200).
16. The regression is weighted by (the inverse of) the number of 11th graders to account for the heteroskedasticity inherent in school-level data.
17. The outcome variable is the percentage of students who did not take the PSAE, which can (and does) equal zero (for 153 of our 573 schools). Thus, we also estimate the relationship between previous test scores and current rates of not taking the test with a Tobit regression. The results are

qualitatively identical. The results from these regressions and for all robustness checks mentioned in footnotes are available from the corresponding author upon request.

18. Including more power terms or dummy variables for the distribution of pass rates yields similar results.
19. The results are essentially unchanged when separate models are estimated for inner-city, suburban, and rural schools, with the exception that the relationship in column (2) is statistically insignificant in suburban schools.
20. See Burtless (1996) for a recent volume discussing education financing.
21. For each model in Table 4, OLS results (not reported) can be used with the IV results to conduct a Hausman test for the consistency of OLS. For each model, the Hausman test rejects the OLS model with a p value under 0.001. A Baseman (1960) test, which assumes at least one of the instruments is valid, is performed to test the validity of the entire set of instruments. These results are included toward the bottom of the table.
22. The Baseman test fails to reject the alternative hypothesis that the set of instruments is invalid.
23. The Baseman test fails to reject the alternative hypothesis that the set of instruments is invalid in the second model. In the third model, however, the Baseman test rejects the set of instruments as being valid.
24. In addition to schools possibly manipulating their scores by influencing who takes the test or by dictating curriculum (i.e., teaching to the test or repetitively requiring practice tests), states have an incentive and opportunity to manipulate the system. One potential pitfall with NCLB is that states, by and large, set their own standards and improvement goals. The standards are subject to the US Department of Education, but the incentive and possibility exists for states to respond to NCLB by lowering standards, not by raising student achievement.

References

Archibold, R. C. (1999) Teachers tell how cheating worked, *New York Times,* 8 December.

Baseman, R. L. (1960) On finite sample distributions of generalized classical linear identifiability test statistics, *Journal of the American Statistical Association,* 55(4), pp. 650–659.

Burtless, G. (Ed.) (1996) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (Washington, DC: Brookings Institution).

Dillon, S. (2003) Playing the standards game, *New York Times,* 25 May.

Dylan, W. (2001) An overview of the relationship between assessment and the curriculum, in: D. E. Scott (Ed.) *Curriculum and Assessment,* pp. 165–182 (Westport, CT: Ablex Publishing).

Figlio, D. N. and Getzler, L. S. (2002) Accountability, ability, and disability: gaming the system, Working Paper #9307, National Bureau of Economic Research, October.

Goodnough, A. (1999) Answers allegedly supplied in effort to raise test scores, *New York Times,* 8 December.

Haney, W. (1993) Minorities and testing, in: M. Fine and L. Weis (Eds) *Silenced Voices: Class, Race, and Gender in the United States Schools,* pp. 45–73 (Albany, NY: State University of Albany Press).

Haney, W. (2000) The myth of the Texas miracle in education, *Education Policy Analysis,* 8(41), pp. 1–79.

Heckman, J. J *et al.* (2002) The performance of performance standards, *Journal of Human Resources,* 37(4), pp. 778–811.

Hillocks, G. (2002) *The Testing Trap: How State Writing Assessments Control Learning* (New York: Teachers College Press).

Hoxby, C. M. (2000) The effects of class size on student achievement: new evidence from population variation, *Quarterly Journal of Economics,* 115(4), pp. 1239–1285.

Illinois State Board of Education (2002) System of support for districts and high priority schools, December.

Jacob, B. A. (2002) Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools, Working Paper #8968, National Bureau of Economic Research, June.

Jacob, B. A. and Levitt, S. D. (2003a) Rotten apples: an investigation of the prevalence and predictors of teacher cheating, Working Paper #9413, National Bureau of Economic Research, January.

Jacob, B. A. and Levitt, S. D. (2003b) Catching cheating teachers: the results of an unusual experiment in implementing theory, in: W. G. Gale and J. Rothenberg-Pack (Eds) *Brookings-Wharton Papers on Urban Affairs,* pp. 185–209 (Washington, DC: Brookings Institution).

Kane, T. J. and Staiger, D. O. (2002) The promise and pitfalls of using imprecise school accountability measures, *Journal of Economic Perspectives,* 16(4), pp. 91–114.

Lewin, T. and Medina, J. (2003) To cut failure rate, schools shed students, *New York Times,* 31 July.

Manski, C. F. (1995) *Identification Problems in the Social Sciences* (Cambridge, MA: Harvard University Press).

Manski, C. F. *et al.* (2002) Using performance standards to evaluate social programs with incomplete outcome data, *Evaluation Review,* 26(4), pp. 355–381.

National Commission on Excellence in Education (1983) *A Nation at Risk: The Imperative for Educational Reform* (Washington, DC: US Government Printing Office).

Pepper, J. V. (2003) Using experiments to evaluate performance standards: what do welfare-to-work demonstrations reveal to welfare reformers?, *Journal of Human Resources,* 38(4), pp. 860–880.

Wilgoren, J. (2001) Possible cheating scandal is investigated in Michigan, *New York Times,* 9 June.