BRIAN A. JACOB
John F. Kennedy School of Government, Harvard University

STEVEN D. LEVITT
University of Chicago and American Bar Foundation

Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory

EDUCATIONAL REFORM is a critical issue in urban areas. Most large urban school districts in the United States suffer from low test scores, high dropout rates, and frequent teacher turnover. Poor performance of city schools induces flight to the suburbs by affluent families with children, eroding the urban tax base. In response to these concerns, the past decade has seen an increasing emphasis on high-stakes testing. While there is evidence such testing has been associated with impressive gains in test scores in some instances, critics have argued that these gains are artificially induced by "teaching to the test."[1] Indeed, much of the observed test score gain has been shown to be test-specific, not generalizing to other standardized tests that seemingly measure the same skills.[2] Even more ominous is the possibility that the emphasis on high-stakes testing induces cheating on the part of students, teachers, and administrators.

We have developed a method for detecting cheating by teachers and administrators on standardized tests.[3] The basic idea underlying the

We would like to thank Marisa de la Torre, Arne Duncan, John Easton, and Jessie Qualls of the Chicago Public Schools for their extensive cooperation on this project. Phil Cook, William Gale, Austan Goolsbee, Janet Pack, and Bruce Sacerdote provided valuable comments on the paper. This paper was completed while the second author was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, California.

1. For testing and gains, see Jacob (2002); Grissmer and others (2000).
2. Jacob (2002); Klein and others (2000).
3. Jacob and Levitt (forthcoming).

method (which is described in greater detail later) is that cheating classrooms will systematically differ from other classrooms along a number of dimensions. For instance, students in cheating classrooms are likely to experience unusually large test score gains in the year of the cheating, followed by unusually small gains or even declines in the following year when the boost attributable to cheating disappears. Just as important as test score fluctuations as an indicator of cheating, however, are telltale patterns of suspicious answer strings—identical blocks of answers for many students in a classroom or cases where students are unable to answer easy questions correctly but do exceptionally well on the most difficult questions. We have concluded that cheating occurs in 3 to 5 percent of elementary school classrooms each year in the Chicago Public Schools (CPS).

Most academic theories, regardless of their inherent merit, fail to influence policy or do so only indirectly and with a long lag. In this paper we report the results of a rare counterexample to this familiar pattern involving collaboration between the CPS and the authors. At the invitation of Arne Duncan, CEO of the Chicago Public Schools, we were granted the opportunity to work with the CPS administration to design and implement auditing and retesting procedures using the tools we developed. With our cheating detection algorithm, we selected roughly 120 classrooms to be retested on the spring 2002 Iowa Test of Basic Skills (ITBS) that was administered to students in the third to eighth grades. The classrooms retested included not only instances suspected of cheating, but also those that had achieved large gains but were not suspected of cheating, as well as a randomly selected control group. As a consequence, the implementation also allowed a prospective test of the validity of the tools we developed.

The results of the retesting provided strong support for the effectiveness of the cheating-detection algorithm. Classrooms suspected of cheating experienced large declines in test scores when retested under controlled conditions. Classrooms not suspected previously of cheating maintained almost all of their gains on the retest. The results of the retests were used to launch investigations of twenty-nine classrooms. While these investigations have not yet been completed, it is expected that disciplinary action will be brought against a substantial number of teachers, test administrators, and principals.

Finally, the data generated by the auditing experiment provided a unique opportunity for evaluating and improving the techniques for detecting cheating. The cheating algorithm was developed without access to multiple observations for the same classrooms. By observing two sets of results from the same classrooms (one from the original test and a second from the retest), we are able for the first time to directly evaluate the predictive power of the various elements of the algorithm. The results suggest improvements to the ad hoc functional form assumptions used in the original research, and also suggest that some of our indicators are much better predictors than are others. By changing the weights used in the algorithm, we should be able substantially to improve the predictive value of the model in future implementations.

In the remainder of the paper we first present background information on teacher cheating and the detection methods. The next section outlines the design and implementation of the retesting procedure. We then report the results of the retests. The final section shows how the data from the retests were used to analyze the predictive value of the various components of the algorithm and identifies a number of possible improvements to the methods.

Teacher Cheating and Its Detection

The emphasis placed on standardized tests in elementary and secondary education has been steadily increasing over the past decade. The recent federal reauthorization of the Elementary and Secondary Education Act (ESEA), which requires states to test students in third through eighth grade each year and to judge the performance of schools based on student achievement scores, is just one prominent example of this trend. Before the passage of that law, every state in the country except Iowa already administered statewide assessment tests to students in elementary and secondary school. Twenty-four states required students to pass an exit examination to graduate from high school. California recently put into place a policy providing for merit pay bonuses of as much as $25,000 for each teacher in schools with large test score gains.

Critics of high-stakes testing argue that linking incentives to performance on standardized tests will lead teachers to minimize other teaching

skills or topics not directly tested on the accountability exam.[4] Studies of districts that have implemented such policies provide mixed evidence, suggesting some improvements in student performance along with indications of increased teaching to the test and shifts away from teaching subjects that were not tested.[5]

A more sinister behavioral distortion is outright cheating on the part of teachers, administrators, and principals, such as erasing student answers and filling in the correct response or telling students the answers.[6] While the idea of elementary school teachers manipulating student answer sheets may seem far-fetched, cheating scandals have appeared in many places, including California, Massachusetts, New York, Texas, and Great Britain.[7] We have provided the first systematic analysis of teacher cheating.[8] We argued that cheating classrooms are likely to share three characteristics: unusually large test score gains for students in the class the year the cheating occurs, unusually small gains the following year for those same students, and distinctive patterns of "suspicious" answer strings.

The first two characteristics relating to test scores are straightforward. Large increases are expected in cheating classrooms because raising test scores is the very reason for the cheating. Unlike gains associated with true learning, however, one expects no persistence in the artificial test score gains due to cheating. Thus if the children in cheating classrooms this year are not in cheating classes next year, one expects the full magnitude of the cheating-related gain to evaporate.

4. Holmstrom and Milgrom (1991).

5. See, for example, Deere and Strayer (2001); Grissmer and others (2000); Heubert and Hauser (1999); Jacob (2001, 2002); Klein and others (2000); Richards and Sheu (1992); Smith and Mickelson (2000); Tepper (2001).

6. As a shorthand, we refer to this behavior simply as teacher cheating, although in using this terminology we are by no means excluding cheating by administrators and principals.

7. For California see Meredith May, "State Fears Cheating by Teachers," *San Francisco Chronicle*, October 4, 2000, p. A1. For Massachusetts, Jon Marcus, "Faking the Grade," *Boston Magazine,* February, 2000. For New York, Loughran and Comiskey (1999). For Texas, Claudia Kolker, "Texas Offers Hard Lessons on School Accountability, *Los Angeles Times*, April 14, 1999, p. 1. For Great Britain, Tony Tysome, "Cheating Purge: Inspectors Out," *Times Higher Education Supplement*, August 19, 1994, p. 1.

8. Jacob and Levitt (forthcoming). In contrast, there is a well-developed literature analyzing student cheating: Aiken (1991); Angoff (1974); Frary, Tideman, and Watts (1977); van der Linden and Sotaridona (2002).

Establishing what factors signify suspicious answer strings is more complicated. Teachers may cheat in a variety of ways. The crudest, most readily detected cheating involves changing answers in a block of consecutive questions so that they are identical for many or all students in a classroom. From the teacher's perspective, this is the quickest and easiest way to alter test forms. A slightly more sophisticated type of cheating involves changing the answers to nonconsecutive questions to avoid conspicuous blocks of identical answers. An even cleverer teacher may change a few answers for each student, but be careful not to change the same questions across students.

We use four separate measures of suspicious strings to detect these varieties of cheating.[9] All four of our indicators are based on deviations by students from the patterns of answers one would expect the students themselves to generate. Thus the first step in analyzing suspicious strings is to estimate the probability each child would give a particular answer on each question. This estimation is done using a multinomial logit framework with past test scores, demographics, and socioeconomic characteristics as explanatory variables. Past test scores, particularly on tests of the same subject, are very powerful predictors of the student answers on a current test.

The first suspicious-string indicator is a measure of how likely it is that, by chance, the single most unusual block of identical answers given by any set of students in the class on any consecutive set of questions would have arisen. This cheating indicator is likely to capture effectively the most naive form of cheating but may not adequately identify more sophisticated types, which are addressed by our second and third measures. The second indicator measures the overall extent of correlation across student answers in a classroom. A high degree of correlation may indicate cheating, since the cheating is likely to take the form of changing haphazardly incorrect answers to shared correct answers. The third indicator captures the cross-question variation in student correlations. If a classroom produces a few questions in which the correlation in student answers is very high but the degree of correlation across students in the classroom on other questions is unremarkable, this suggests intervention on the part of the teacher on the questions for which answers are highly

9. For the formal mathematical derivation of how each of the cheating indicators is constructed, see Jacob and Levitt (forthcoming).

correlated. The fourth and final indicator of a suspicious string measures the extent to which students in a classroom get the easy questions wrong and the hard questions correct. In other words, by comparing the responses given by a particular student to those of all other students who got the same number of correct answers on that test, we are able to construct an index of dissimilarity in the answers each student gives.

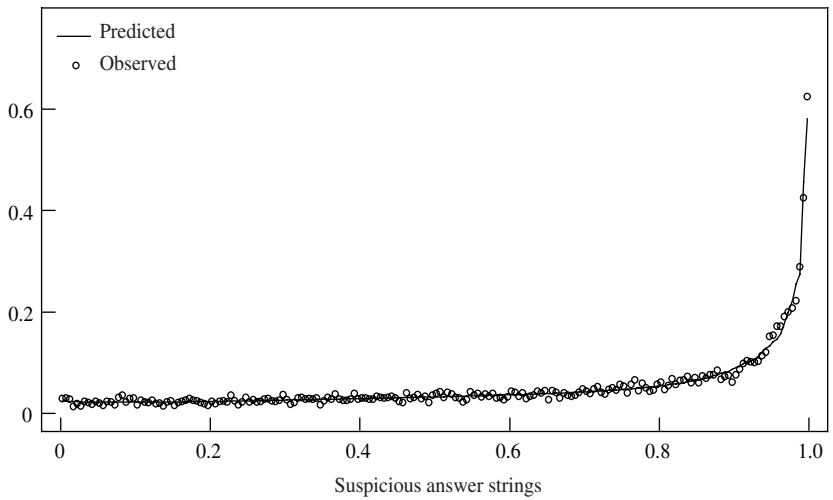
To construct an overall summary statistic measuring the degree of suspiciousness of a classroom's answers, we rank the classes from least to most suspicious within subject and grade on each of the four measures. We then take the sum of squared ranks as our summary statistic. By squaring these ranks, greater emphasis is put on variations in rank in the right-hand tail (that is, the most suspicious part) of the distribution. A parallel statistic is constructed for the two test-score-gain measures corresponding to a given year's gain and the following year's gain for students in the class.

Although skepticism about the ability of these indicators to identify cheating might seem warranted, we present a range of evidence supporting the argument that these measures have predictive power empirically.[10] For instance, among classrooms that have large test score gains this year, children in classrooms that have suspicious answer strings do much worse on standardized tests the following year. This suggests that big test score gains that are not accompanied by suspicious answer strings represent real learning (which partially persists to the following year), whereas large test score gains accompanied by suspicious strings are likely due to cheating. Second, there tend to be strong correlations across subjects within a classroom and within classrooms over time in the incidence of our cheating indicators. That result is consistent with a subset of teachers who tend to cheat repeatedly. Third, the apparent cheating is highly correlated with certain incentives. For example, cheating is more likely to occur in poorly achieving schools that face the risk of being put on probation, and when social promotion is ended, cheating increases in the affected grades.

Perhaps the most convincing evidence of the usefulness of the cheating indicators, however, is visual. In figure 1 the horizontal axis reflects how suspicious the answer strings are in a classroom and the vertical axis is the probability that students in a classroom experience an unusually

10. Jacob and Levitt (forthcoming).

Figure 1. The Relationship between Unusual Test Scores and Suspicious Answer Strings



Notes: The measure of suspicious answer strings on the horizontal axis is measured in terms of the classroom's rank within its grade, subject, and year, with 0 representing the least suspicious classroom and 1 representing the most suspicious classroom. The 95th percentile is the cutoff for both the suspicious answer strings and test score fluctuation measures. The results are not sensitive to the cutoff used. The observed points represent averages from 200 equally spaced cells along the x-axis. The predicted line is based on a probit model estimated with seventh order polynomials in the suspicious string measure.

large test score gain in the current year followed by an unexpectedly small increase (or even a decline) in the following year.[11] Up to roughly the 90th percentile on suspicious strings and even higher, there is little or no relationship between the frequency of large test score fluctuations and suspicious strings in this subset of the data. Based on these data, if one were to predict what the pattern in the rest of the data would likely be, a continued flat line might be a reasonable conjecture. In actuality, however, there is a dramatic spike in the frequency of large test score fluctuations for classrooms that have very suspicious answers, as evidenced in the right-hand tail of figure 1.

Our interpretation of this striking pattern is that the enormous increase in unexpected test score fluctuations in the right-hand side of the figure

11. More precisely, to qualify as having large test score fluctuations in this figure, a classroom must be in the top 5 percent of classrooms with respect to the magnitude of the current year's increase relative to the following year's decrease.

reflects the fact that teacher cheating increases the likelihood of suspicious strings and of large test score increases. In a previous study we formally demonstrated that under carefully articulated assumptions, the area beneath the curve in the right-hand tail of figure 1 measures the overall incidence of teacher cheating. Empirically, our findings imply that as many as 5 percent of the classrooms in the CPS show evidence of cheating on the ITBS in any given year.

Using the Cheating Detection Algorithm in 2002 ITBS Testing

Each spring, 100,000 CPS students take the ITBS test. The results of this test determine which schools will be placed on academic probation or reconstituted, which students will be required to attend summer school and potentially be retained (third, sixth, and eighth grade only), and which students are eligible to apply to the most sought-after test-based magnet high schools in the CPS system (seventh grade).

The accountability department of CPS conducts retests of the ITBS in about 100 classrooms annually to achieve quality assurance. The retests, which use a different version of the exam, occur three to four weeks after the initial testing. Specially trained staff in the accountability office administer the retests. Unlike the initial round of testing, which is subject to relatively lax oversight and control and potentially affords various school staff members access to the test booklets, the retest answer sheets are closely guarded. Until the past few years, classrooms were randomly selected for retests.[12] Since then, retests have been focused on those classrooms achieving the largest test score gains relative to the previous year. Formal investigations have been undertaken when major discrepancies arise between the official testing and the retesting, but punishment is rare. We are aware of only one instance in the last decade in which disciplinary actions have been taken in CPS as a consequence of teacher cheating on ITBS.

In spring 2002 Arne Duncan, CEO of the CPS, having read our earlier work on teacher cheating, invited us to work with the staff of CPS in selecting the classrooms to be retested. The only real constraint on the

12. The exception to this rule was that if credible accusations of cheating were made about a classroom, that classroom would be retested with certainty.

implementation of the audits was that budget limitations restricted the total number of classrooms audited to no more than 120; our earlier research on cheating estimated that roughly 200 classrooms cheated each year in CPS. Thus the budget constraint meant that we were able to audit only a small fraction of the suspected cheaters.

Selecting individual classrooms with the goal of *prospectively* identifying cheating raised an important issue because the original method to detect cheating we developed in earlier work relies heavily on the availability of the following year's test scores (to determine whether large test score gains in the current year are purely transitory as would be suspected with cheating).[13] In selecting classrooms to retest, however, the next year's test scores did not, of course, exist. As a consequence, the choice of classes to audit could depend only on test scores from the current and previous years, as well as suspicious answer strings from the current year.

Table 1 outlines the structure of the design we developed. Classrooms to be audited were divided into five categories. The first group exhibited both unusually large test score gains and highly suspicious patterns of answer strings. These classrooms were judged to be the most likely to have experience cheating. A second group had very suspicious patterns of answer strings but did not have unusually large test score gains. Those patterns are consistent with a bad teacher who has failed to teach the students adequately and has attempted to cover up this fact by cheating. Thus these classrooms were suspected of high rates of cheating. A third category of classrooms encompassed those for which anonymous allegations of cheating had been made to CPS officials. There were only four of these, none of which would have otherwise made the cutoff for inclusion in our first two groups of suspected cheaters.

The remaining two categories of classrooms audited were not suspected of cheating; they served as control groups. One category included classrooms with large test score gains but with answer string patterns that did not point to cheating. These classrooms were judged likely to have good teachers capable of generating big test score gains without resorting to devious means.[14] As such, they provided an important comparison with the suspected cheaters with large gains. A fifth and final set had classrooms

13. Jacob and Levitt (forthcoming).

14. Alternatively, these classes may have had cheating, but of a form that our methods failed to detect.

Table 1. Design of the 2002 Sample of Classrooms to Be Audited[a]

Category of classroom	*Comments*	*Did the classroom have suspicious patterns of answer strings on spring 2002 ITBS?*	*Did the students in the classroom achieve unusually high test score gains between 2001 and 2002 ITBS?*	*Prediction about how test scores will change between spring 2002 ITBS and audit test*	*Number of classrooms audited*
Suspected cheaters					
Most likely cheaters	Look suspicious on both dimensions	Yes	Yes	Big decline in test scores when audited	51
Bad teachers suspected of cheating	Even though they cheat, test score gains not that great because teach students so little	Yes	No	Big decline in test scores when audited	21
Anonymous tips	Complaints to CPS	Varies	Varies	Big decline in test scores if complaint is legitimate	4
Control groups					
Good teachers	Big gains, but no suspicion of cheating	No	Yes	Little change between original test and audit	17
Randomly selected rooms	Control group	No	No	Little change between original test and audit	24

a. Not all classrooms were administered both reading and mathematics tests. In particular, to conserve resources, each classroom in the randomly selected control group was given only one portion of the test (either reading, or one of the three sections of mathematics). For the other classrooms, either the entire test was administered, just reading, or all three sections of the mathematics exam.

that were randomly chosen from all remaining classrooms. These classrooms were also unlikely to have high rates of cheating.

With the exception of those attributed to anonymous tips and the classrooms that were randomly chosen, we did not employ a rigid cutoff rule for allocating classrooms into the various categories. To be assigned to the first or second category, a classroom generally needed to be in the top few percent of those with suspicious answer strings on at least one subject test. For the first category the classroom also typically had to be in the top few percent on test score gains. In cases where multiple subject tests had elevated levels of suspicious strings, the cutoffs were sometimes relaxed. In addition, some classrooms that appeared suspicious but otherwise would not have made it into categories one or two, were included because other classrooms in the same school did qualify and we were interested in isolating schoolwide instances of cheating.

Dividing classrooms to be audited in this manner provides two benefits. First, the presence of two control groups (the randomly selected classrooms and the rooms that showed large achievement gains but did not have suspicious answer strings) allows a stronger test of the hypothesis that other classrooms are cheating. In the absence of these control groups, one might argue that large declines in the retest scores relative to the initial test in classrooms suspected of cheating are due to reduced effort on the part of students on the retest.[15] By isolating a set of classrooms that made large gains in achievement but did not appear to cheat, we are able to determine the extent to which declines in scores among the high-achieving suspected cheaters may simply be the consequence of reversion to mean. Second, including the control groups allows us to more effectively test how various components of our model are working in identifying cheating after it has occurred. The cost of the retest structure with the inclusion of a control group meant that we were able to retest fewer classrooms suspected of cheating. Of the 117 retested, 76 were suspected of cheating (51 with suspicious strings and large test score gains, 21 with suspicious strings but no large gains in scores, and 4 identified by anonymous tips). Again, there were many more classrooms

15. Indeed, when administering the retest, the proctors are told to emphasize that the outcome of the retest will not affect the students in any way. These retests are not used to determine summer school or magnet school eligibility and are not recorded in a student's master file.

that looked equally suspicious or nearly so but were not retested because of the budget constraints.[16]

In some cases, classrooms were retested on only the mathematics or the reading tests, not both.[17] In particular, those that were suspected of cheating on only the mathematics test were generally not retested on reading. Classes for which there were anonymous tips were retested only on reading. Finally, in the randomly selected control group, either the mathematics or the reading test was administered, but not both. In the results presented in the next section, we report test score comparisons only for those subjects on which retesting took place.

Results of the Retests

The basic results of the retests are shown in table 2. For most of the categories of classrooms we defined, six average test score gains are presented (three each for mathematics and reading).[18] For the randomly selected classrooms, there were so few data that we lumped together mathematics and reading. For the classes identified by anonymous tips, audits took place only on reading tests, so we do not report mathematics scores. In all cases the test score gains are reported in terms of standard score units, the preferred metric of the CPS. A typical student gains approximately fifteen standard score units in an academic year.

The first three columns show the results on the reading test (and the combined reading and mathematics test results for the randomly selected classrooms). Column 1 presents test scores between the spring 2001 and spring 2002 ITBS (the actual test, not the retest). For all classrooms in

16. Aware of the overall resource constraints, we provided an initial list of classrooms to CPS that had 68, 36, and 25 classrooms in categories 1, 2, and 4 respectively. Had resources been unlimited, more suspected classrooms could have been identified. Within each category, classrooms on our list were not ordered by degree of suspicion. The choice of which schools to retest from our list was made by CPS staff. In response to resistance on the part of principals at heavily targeted schools, a limited number of classrooms were retested at any one school. In a few cases, principals and parents simply refused to allow the retests to be carried out.

17. The mathematics portion of the ITBS has three sections. Every class retested on mathematics was given all three sections of the exam, even if the classroom was suspected of cheating on only one or two sections of the initial test.

18. When we talk about test score gains, we are referring to the change in test scores for a given student on tests taken at different times.

Table 2. Results for Spring 2002 ITBS and 2002 Audit Test[a]

Standard score units

	Reading gains			*Mathematics gains*		
Category of classroom	*Spring 2001 to spring 2002*	*Spring 2002 and 2002 retest*	*Spring 2001 and 2002 retest*	*Spring 2001 to spring 2002*	*Spring 2002 and 2002 retest*	*Spring 2001 and 2002 retest*
All classrooms in CPS	14.3	. . .	. . .	16.9	. . .	. . .
Most likely cheaters (N = 36 on math, N = 39 on reading)	28.8	–16.2	12.6	30.0	–10.7	19.3
Bad teachers suspected of cheating (N = 16 on math, N = 20 on reading)	16.6	–8.8	7.8	17.3	–10.5	6.8
Anonymous tips (N = 0 on math, N = 4 on reading)	26.2	–6.8	19.4	. . .	. . .	. . .
Good teachers (N = 17 on math, N = 17 on reading)	20.6	0.5	21.1	28.8	–3.3	25.5
Randomly selected classrooms (N = 24 overall, but only one test per classroom)	14.5	–2.3	12.2	14.5	–2.3	12.2

a. Because of limited data, mathematics and reading results for the randomly selected classrooms are combined. Only data for the first two columns are available for all CPS classrooms because audits were performed only on a subset of classrooms.

CPS (those that are retested and those that are not), the average gain on the reading test was 14.3 standard score points. Classrooms identified in advance as most suspicious achieved gains almost twice as large; that is, students in these classes tested roughly two grade equivalents higher than they had in tests in 2001. The control group of good teachers achieved gains that were large (20.6) but not as great as those of the suspected cheaters. Bad teachers suspected of cheating had test score gains slightly above the average CPS classroom. The scores of the randomly selected classes were in line with the scores of the CPS average, as would be expected.

Column 2 shows how the reading test scores changed between the spring 2002 test and the spring 2002 retest conducted a few weeks later. The results are striking. The most likely cheaters saw a decline of 16.2 standard score points, or more than a full grade equivalent. The bad teachers suspected of cheating saw declines of 8.8 standard score points. The classes identified by anonymous tips lost 6.8 points. In stark contrast the classrooms with good teachers actually registered small *increases* on the audit test relative to the original.[19] The randomly selected classrooms lost 2.3 points, or only one-seventh as much as the most likely cheaters. The fact that the two control groups (those with good teachers and the randomly selected classes) saw only small declines suggests that the impact of decreased effort by students on the retest is likely to be minimal. The much larger decline in scores on the audit test for the suspected cheaters is consistent with the hypothesis that their initial reading scores were inflated by cheating.

Column 3 shows the gain in test scores between the spring 2001 ITBS and the spring 2002 retest and thus represents an estimate of the "true" gain in test scores, once the 2002 cheating is eliminated (the figures in column 3 are simply the sums of those in columns 1 and 2).[20] The largest "true" gains, as would be expected, are in the classrooms identified as having good teachers. The classes most likely cheating that scored so high on the initial test look merely average in terms of "true" gains, sug-

19. As noted earlier, mathematics and reading scores are lumped together for the randomly selected classrooms, so the decline of 2.3 reported in column 2 would be applicable here as well.

20. This statement is subject to the caveat that effort might have been weaker on the retest and that the spring 2001 scores might themselves be inflated by cheating that occurred in the previous year.

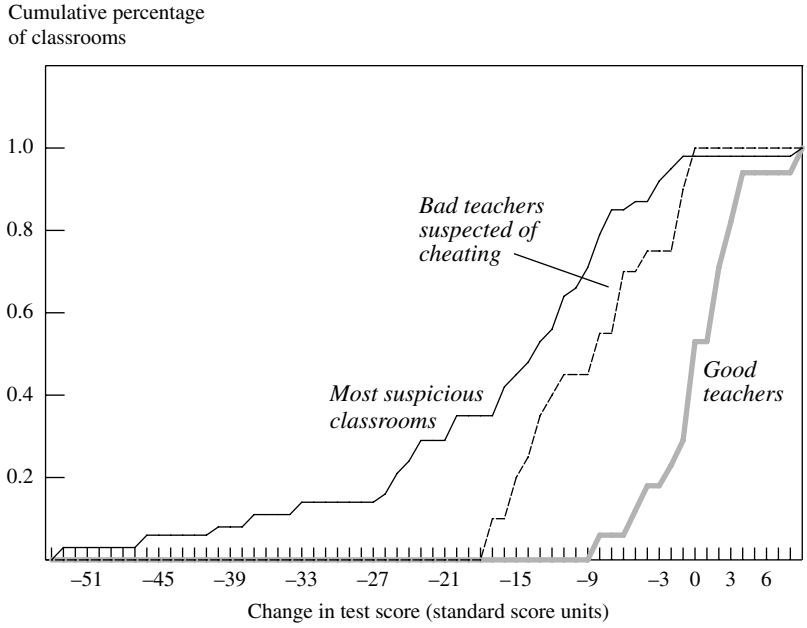
gesting that all of their apparent success is attributable to cheating. For the bad-teacher category, once the cheating is stripped away, the reading performance is truly dismal: gains of just 7.8 standard score points, or little more than half a grade equivalent in a year. Classrooms identified through anonymous tips experienced some declines on the retest but continued to score well above average.

Columns 4 through 6 show results parallel to those in the first three columns, but for the mathematics tests. The results are generally similar to those for reading, but less stark.[21] The good teachers have baseline mathematics gains commensurate with those of the most likely cheaters (column 4), which was not true in reading. The results of the audit tests in column 5 once again show large declines for the two categories of classrooms suspected of cheating (declines of more than 10 standard score points in each case). The classrooms with good teachers also show a small decline in mathematics scores on the retest (3.3 points), unlike on the reading retest, where they gained. Finally, in column 6 a notable difference between the results for reading and mathematics is that the classrooms considered most likely to be cheating showed above average "true" gains in mathematics, which was not the case for reading. This result is likely due to the fact that the modified algorithm used for predicting cheaters relies in part on large test score gains and thus is biased toward identifying classrooms that have large real gains. (In contrast, the retrospective algorithm used to assess teacher cheating in our earlier published work is specifically designed to be neutral in this regard; without access to the next year's test scores, however, this neutrality is lost). In other words, the false positives generated by the prospective algorithm are likely to be concentrated among classrooms with large true gains.[22]

21. A partial explanation for why the results on the mathematics test are less stark than those for reading is that the mathematics test is made up of three parts, unlike the reading test, which is in one self-contained section. When the retests were conducted, classrooms suspected of cheating on any of the three mathematics sections were retested on the entire test. Thus, included in the mathematics results are some classes where there was strong evidence of cheating on one part of the exam but not on another part. Even when the results are further disaggregated, identifying particular sections of the exam where classes were judged beforehand as likely to have cheated, the results are not as clean as for the reading test.

22. Alternatively, it could just be that good teachers are also more likely to cheat. We are skeptical of this hypothesis since using our retrospective measure, we have found cheating to be concentrated in the lowest-achieving schools and classrooms; Jacob and Levitt (forthcoming).

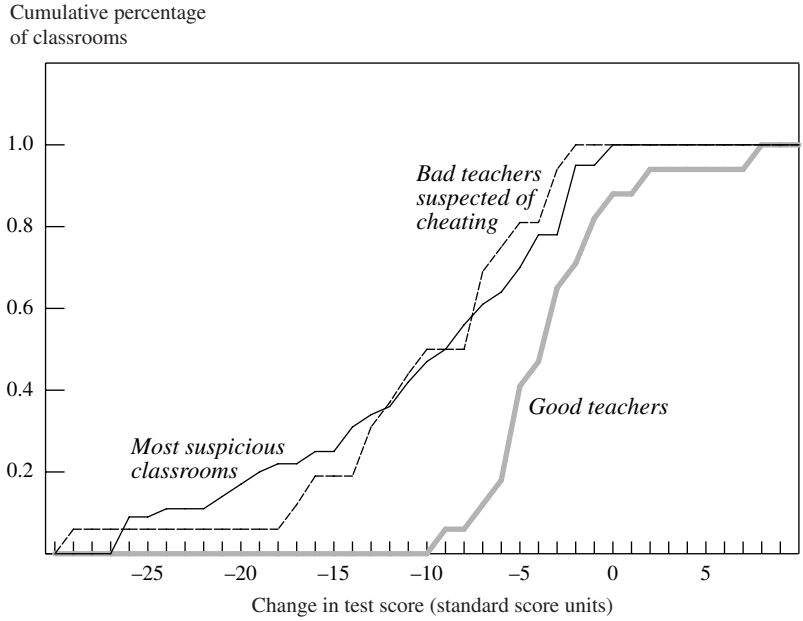
Figure 2. Cumulative Distribution of Change in Reading Test Scores between Initial Test and Retest, by Audit Category



Figures 2 and 3 show the cumulative distribution of changes in reading and mathematics test scores between the initial spring 2002 test and the retest for classrooms in different categories. These figures highlight the stark differences between the classes predicted to be cheating and those identified as having good teachers. The vertical axis is the cumulative percentage of classrooms with a test score change between the initial test and the audit that is less than the value measured on the horizontal axis. Three cumulative distributions are plotted in each figure, corresponding to the classrooms previously considered most suspicious—those with bad teachers suspected of cheating—and those with good teachers. The striking feature of the figure is how little overlap there is between the distributions of the cheating and the good teachers.

In figure 2 the worst outcome for the most suspicious classrooms was a decline of 54 points (roughly three grade equivalents). Many classes in this category experienced very large losses. The bad teachers suspected of cheating are not represented by a long left tail like the most suspicious cheaters, but have a high concentration of cases in which there are dou-

Figure 3. Cumulative Distribution of Change in Mathematics Test Scores between Initial Test and Retest, by Audit Category



ble-digit losses. In contrast, the single biggest test score decline experienced by a good-teacher classroom on reading is 7 standard points (as indicated by the cumulative distribution rising above zero at that point for the good-teacher curve). More than 80 percent of the most suspicious classrooms experienced losses greater than that, and almost 60 percent of classrooms with bad teachers saw bigger declines. About one-third of the good-teacher classrooms experienced test score gains, whereas virtually none of the suspected cheating classrooms did.

The results shown in figure 3 are similar. The primary differences between the two figures are, first, that the distribution of outcomes for the most suspicious teachers and the bad teachers suspected of cheating are almost identical on the mathematics test and, second, the gap between the good teachers and the suspected cheaters is not quite as pronounced. Figures 2 and 3 demonstrate that the differences in means presented in table 2 are not driven by a few outliers, but rather represent systematic differences throughout the entire distribution. One implication of these findings is that our methods not only provide a means of identifying

potential cheating classrooms, but also that they are at least as successful in identifying classrooms with good teachers whose students' gains are legitimate and who are possibly deserving of rewards and of analysis as exemplifying the best instructional practices.[23]

Thus far we have focused exclusively on the classroom as the unit of analysis. Another question of interest is the extent to which cheating tends to be clustered in particular schools, and if so, why?[24] Unfortunately, the way in which the audits were conducted limits the amount of light we are able to shed on answers to this question. The CPS officials who determined which classrooms to audit intentionally tried to avoid retesting large numbers of classes in particular schools because such an action could elicit negative reactions. There are at least two schools, however, in which the audits provided systematic evidence of centralized cheating likely to have been perpetrated by school administrators. These cases are currently under investigation by CPS. More generally, however, it appears that most cheating incidents are consistent with cheating by teachers rather than by administrators.

Using the Retests to Evaluate and Improve Cheating Detection

This paper has so far focused on evaluating how effective the methods previously developed were in identifying prospective cheaters. The retest also provides a unique opportunity for refining the cheating detection algorithm. In developing the algorithm we made a number of relatively arbitrary functional form and weighting assumptions, which can be tested using the data generated by the retests.

Our measure of how suspicious a classroom's answer strings are is based on an average of that class's rank on each of the four indicators

23. Some caution, however, must be exercised in discussing "good" teachers. Our findings suggest that classrooms with large gains in test scores that do not have suspicious patterns of answer strings can maintain their gains on retests. Whether the large test score gains are the result of artificially low test scores in the previous year (due perhaps to a bad teacher or adverse test conditions in that year) is not something we have explored.

24. Possible explanations include cheating by central administration, explicit collusion by corrupt teachers (teachers generally do not proctor their own students during the exam, so cooperation of other teachers aids in cheating), a school environment or culture that encourages cheating, or systematic differences in incentives among schools (for example, because schools performing badly are threatened with probation and reconstitution).

discussed earlier. The indicators have been given equal weight in the algorithm. Moreover, although greater weight is given to variation in the right-hand tail of the distribution of each measure, the weighting function (squaring the ranks) used was chosen somewhat arbitrarily. Using the results of the retest, we are able to test the validity of these assumptions by estimating regressions of the form

$$Change\_in\_test\_score_{cs} = Suspicious\_string\_measures_{cs}'\Gamma + \gamma_s + \theta_{g,},$$

where the left-hand-side variable is the change in test score between the initial spring 2002 test and the audit for a given classroom c on subject s. The primary right-hand-side variables are the suspicious string measures, which will be entered in a variety of ways to test the predictive ability of alternative functional form and weighting assumptions. The unit of observation in the regression is a classroom-subject test. Subject- and grade-fixed effects are included in all specifications. The four subject tests (reading comprehension and three mathematics tests) are pooled together and estimated jointly. In some cases we also include the gain between the spring 2001 and spring 2002 ITBS tests as a control for possible mean reversion on the retest. The suspiciousness of a classroom's answers on other subject tests on the same exams is also sometimes included as a covariate in the model. The standard errors are clustered at the classroom level to account for within-classroom correlation across different exams.

It is important to note that the sample of classrooms for which we have retest data (and thus can estimate the equation) is a highly selective one in which extreme values of suspicious answer strings are greatly overrepresented. On the one hand, this is desirable because the parameters are being identified from the part of the distribution that has many cheaters. On the other hand, it is possible that the inference from this select sample will be misleading if applied out of sample to the whole set of classrooms. When thinking about how to improve our algorithm's prospective ability to identify cheaters, that latter (potentially misleading) exercise is precisely what we have in mind. So some caution is warranted.

The first column of table 3 presents the results using the overall measure of suspicious strings that we developed in our initial paper. To aid in interpretation, we use a simple framework in which two indicator vari-

Table 3. Suspicious Answer Strings and Score Declines on the 2002 Retest[a]

Standard score units

Measures of suspicious answer strings	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Overall measures (omitted category is 1–89th percentile)								
Class is in 99th percentile on overall measure	–14.2 (1.6)	. . .	–6.0 (2.6)	–5.5 (2.6)	–11.9 (1.6)	. . .	–5.6 (2.5)	–5.1 (2.5)
Class is in 90th–98th percentile on overall measure	–4.3 (0.9)	. . .	–1.4 (1.4)	–1.0 (1.4)	–3.6 (1.0)	. . .	–1.3 (1.3)	–0.8 (1.3)
Number of individual measures on which class is in 99th percentile (omitted category is zero)								
Four	. . .	–21.1 (2.3)	–14.9 (3.5)	–12.4 (3.3)	. . .	–18.0 (2.0)	–12.2 (3.4)	–9.3 (3.0)
Three	. . .	–11.3 (2.2)	–7.6 (2.6)	–6.0 (2.5)	. . .	–9.0 (2.1)	–5.5 (2.5)	–3.7 (2.5)
Two	. . .	–8.1 (2.2)	–5.2 (2.2)	–4.5 (2.1)	. . .	–6.5 (2.0)	–3.8 (2.0)	–3.0 (2.0)
One	. . .	–4.3 (1.2)	–2.4 (1.4)	–2.3 (1.3)	. . .	–3.7 (1.3)	–1.9 (1.4)	–1.8 (1.4)
Number of individual measures on which class is in 90th–98th percentile (omitted category is zero)								
Four	. . .	–8.7 (2.1)	–4.1 (3.3)	–2.5 (3.4)	. . .	–7.6 (2.1)	–3.2 (3.1)	–1.4 (3.1)
Three	. . .	–4.1 (1.2)	–1.8 (1.7)	–0.7 (1.6)	. . .	–3.2 (1.2)	–1.1 (1.8)	0.3 (1.7)
Two	. . .	–5.4 (1.4)	–3.9 (1.8)	–3.0 (1.7)	. . .	–5.1 (1.4)	–3.7 (1.7)	–2.6 (1.6)
One	. . .	–4.6 (1.1)	–3.9 (1.2)	–3.0 (1.2)	. . .	–4.8 (1.0)	–4.2 (1.1)	–3.0 (1.1)
Average number of categories in 99th percentile on *other* subjects	. . .	. . .	. . .	–1.3 (0.6)	. . .	. . .	. . .	–1.5 (0.6)
Average number of categories in 90th–98th percentile on other subjects	. . .	. . .	. . .	–0.8 (0.5)	. . .	. . .	. . .	–1.0 (0.6)
Test score gain, spring 2001 to spring 2002	. . .	. . .	. . .	. . .	–.24 (.06)	–.22 (.06)	–.21 (.06)	–.22 (.05)
Summary statistic								
R^2	.462	.518	.530		.512	.559	.569	.582

a. The dependent variable is the change in the mean standard score between the spring 2002 ITBS and the retest for students taking both exams. The sample is the classrooms that were retested in spring 2002. The unit of observation is a classroom subject. Sample size is 316. Grade-fixed effects and subject-fixed effects are included in all regressions. Standard errors are in parantheses and are clustered to take into account correlation within classrooms across different subject tests.

ables correspond to whether a classroom is in the 99th percentile on this measure or between the 90th and 99th percentiles. We have experimented with a fuller parameterization, but this sparse specification appears to capture the relevant variation adequately. Classrooms in the 99th percentile on the overall measure of suspicious strings on average lose 14.2 standard score points (about one grade equivalent) on the retest relative to the omitted category (classes below the 90th percentile). This result is statistically highly significant. Classes in the 90th to 98th percentiles lose only one-third as much, although the result is still statistically significant.[25] Thus there appears to be a sharp discontinuity occurring in the last 1 percent of the distribution. In the sample used to estimate this regression, we can explain almost half of the variation in the retest results using these two variables alone.

Column 2 adopts a different functional form for the measure of suspicious answer strings. Rather than aggregating over the four indicators, we count the number of individual indicators for which a classroom is in the 99th percentile, or alternatively, the 90th percentile. Relative to the first column, the second column emphasizes classrooms that look extreme on particular measures (although possibly not extreme at all on others) relative to classrooms that are somewhat elevated on all four measures. Being in the 99th percentile on all four measures individually—an extreme outcome—is associated with a decline of 21.1 points on the retest relative to the omitted category, which is below the 90th percentile on all four measures. Although there is a large difference between being in the 99th percentile on all four measures rather than on three of four (–21.1 compared to –11.3), the marginal impact of an extra indicator above the 99th percentile is about 4 standard score points otherwise. Having one test score above the 90th percentile (but below the 99th) is associated with as great a decline in test scores as having one test above the 99th percentile, but there is no incremental impact of having two or three measures above the 90th percentile. The explanatory power of the specification is substantially higher than that of the first column, although this is in part due to the greater degrees of freedom in the model.

25. If one allows the impact of the 90th to 94th percentile to differ from the 95th to 98th, one cannot deny that the coefficients are identical on those two variables. Indeed, the point estimate on the 90th to 94th percentiles is slightly larger than that on the 95th to 98th.

Further evidence of the usefulness of including the additional detail provided by the model in column 2 is presented in column 3, which nests the models of the preceding two specifications. The coefficients on the aggregate measure in the first two rows fall to less than half their previous magnitude, and only for the 99th percentile variable is the estimate statistically different from zero. In contrast, the indicator variables for the separate measures continue to enter strongly and with a pattern similar to the one before. The R^2 of the nested model in column 3 is only slightly higher than that of column 2. These results suggest that our initial approach to aggregating the information in the original paper (along the lines of column 1) is less effective in predicting outcomes than the alternative presented in column 2.

When the suspiciousness of answer strings on other parts of the exam is added to the specification (column 4), the results are not greatly affected. Observing suspicious answers on the remainder of the test is predictive of greater test declines on the audit, although the magnitude of the effect is relatively small. Even having all four indicators above the 99th percentile on all three of the other subject tests (compared to none of the indicators above the 90th percentile on any of the other subjects) is associated with only a 5 point test score decline on the audit. Thus, while pooling information across subject areas is somewhat useful in identifying cheating, it is much less potent than the information contained in the answer strings to the actual subject test.

Columns 5 to 8 replicate the specifications of the first four columns, but with the baseline test score gain from spring 2001 to spring 2002 included as a regressor. In most cases the results are somewhat attenuated by the inclusion of this variable, which enters significantly negative with a coefficient of roughly –.20. The general conclusions, however, are unaltered.[26]

The specifications in table 3 give equal treatment to each of the four suspicious string measures. Table 4 relaxes that constraint, allowing separate coefficients on each of the measures. Columns 1 and 3 include only indicator variables for being in the 99th percentile on the different measures; columns 2 and 4 also include dummies for the 90th to 98th per-

26. We are guarded in our interpretation of this coefficient and these specifications in general, however, because in results not presented in the table we obtain a coefficient close to zero on this mean reversion variable when we limit the sample to classrooms not suspected of cheating (that is, good teachers and randomly selected controls).

Table 4. Performance of the Individual Suspicious String Indicators in Predicting Score Declines on the Retest[a]

Standard score units

Cheating indicator	*1*	*2*	*3*	*4*
Hard questions right, easy questions wrong				
99th percentile	–9.6	–10.4	–8.7	–9.5
	(2.0)	(2.0)	(1.7)	(1.7)
90th–98th percentile	. . .	–3.6	. . .	–3.1
		(1.2)		(1.2)
Identical answer blocks				
99th percentile	–3.0	–3.9	–1.2	–1.9
	(1.8)	(1.9)	(1.8)	(2.0)
90th–98th percentile	. . .	–2.5	. . .	–1.7
		(1.1)		(1.0)
High overall correlation across students				
99th percentile	–5.3	–5.7	–4.4	–4.9
	(2.2)	(2.4)	(1.9)	(2.1)
90th–98th percentile	. . .	–1.8	. . .	–1.6
		(1.0)		(1.0)
High variance in correlation across questions				
99th percentile	–1.8	–0.8	–2.4	–1.7
	(2.5)	(2.6)	(2.3)	(2.4)
90th–98th percentile	. . .	0.6	. . .	0.3
		(1.1)		(1.1)
Test score gain, spring 2001 to spring 2002	. . .	. . .	–.23	–.20
			(.06)	(.06)
Summary statistic				
R^2	.482	.524	.529	.558

a. The dependent variable is the change in the mean standard score between the spring 2002 ITBS and the retest, for students taking both exams. The sample is the classrooms that were retested in spring 2002. The unit of observation is a classroom subject. Sample size is 316. Grade-fixed effects and subject-fixed effects are included in all regressions. Standard errors are in parentheses and are clustered to take into account correlation within classrooms across different subject tests.

centiles. The final two columns allow for mean reversion. The striking result is that being in the 99th percentile on our measure of students getting the hard questions right but the easy questions wrong is much more effective in predicting score declines on the retest than are the other three measures. The implied decline of roughly 10 standard score points associated with being above this threshold is about the same magnitude as being in the 99th percentile on all three of the other measures. The second most effective cheating indicator is a high degree of overall correlation across student answers. Perhaps surprisingly, identical blocks of answers, which are so visually persuasive, are not particularly good predictors of declines on the retest. This measure is only borderline statisti-

cally significant, and one cannot reject equality of coefficients between being in the 99th percentile and the 90th to 98th percentiles. A large variance in the extent of correlation across questions on the test is the worst predictor among the four measures. None of the coefficients on this indicator are statistically significant and all of the point estimates are small in magnitude.

The results of table 4 suggest that our initial formulation of the suspicious string measures, which used equal weights for all four indicators, would be improved by placing greater emphasis on the measure reflecting students' getting the hard questions right and the easy ones wrong, and by deemphasizing or eliminating altogether the measure of variance across questions.

Conclusions

This paper summarizes the results of a unique policy implementation that allowed a test of tools for predicting and detecting cheating that we had developed. The results of retests generally support the validity of these tools for identifying teacher cheating. Classrooms selected as likely cheaters experienced dramatic declines in scores on retests, whereas classes identified as having good teachers and the randomly selected classrooms experienced little or no decline. In addition, the availability of the retest data provided a direct test of the methods developed, yielding important improvements in the functional form and weighting assumptions underlying the algorithm, which should make the algorithm even more effective in future applications.

On a more practical level, the implementation demonstrated the value of these tools to school districts interested in catching cheaters or deterring future cheating. Out of almost 7,000 potential classrooms, our methods isolated approximately 70 suspicious classrooms that were retested (as well as many more equally suspicious classrooms that were not retested because of budget constraints). Of these 70, almost all experienced substantial declines on the retest, which indicated cheating. In 29 classrooms the declines in test scores were particularly great (more than one grade-equivalent on average across the subjects retested). CPS staff undertook further investigation of these 29 classrooms, including analysis of erasure patterns and on-site investigations. Although disciplinary

actions are still in progress at the time of this writing, there is every indication that for the first time in recent history, a substantial number of cheating teachers will be disciplined for their actions. If punishment is indeed handed out, then estimating the deterrent effect of this punishment on cheating on next year's test will be a potentially interesting subject for exploration.

Although our primary focus has been on the negative outcome of cheating, the positive aspect of this algorithm also deserves emphasis. Using these tools, we were able to identify a set of classrooms that made extraordinary test score gains without any indication of cheating. Without our tools, distinguishing between cheaters and outstanding teachers posed a difficult task. Consequently, identifying outstanding teachers was a tricky endeavor. With our algorithm, however, we can be almost certain that classrooms that do not have suspicious answer strings were not cheating (at least not in ways that lead to test score declines on retests), allowing for a system of rewards that will not inadvertently be directed toward cheaters.

Explicit cheating of the type we identify is not likely to be a serious enough problem by itself to call into question high-stakes testing, both because it is relatively rare (only 1 to 2 percent of classrooms on any given exam) and likely to become much less prevalent with the introduction of proper safeguards such as the cheating detection techniques we have developed. However, our work on cheating highlights the nearly unlimited capacity of human beings to distort behavior in response to incentives. The sort of cheating we catch is just one of many potential behavioral responses to high-stakes testing. Other responses, like teaching to the test and cheating in a subtler manner, such as giving the students extra time, are presumably also present but are harder to measure. Ultimately, the aim of public policy should be to design rules and incentives that provide the most favorable trade-off between the real benefits of high-stakes testing and the real costs associated with behavioral distortions aimed at artificially gaming the standard.

Comments

Philip J. Cook: W. C. Fields, acting in the movie *You Can't Cheat an Honest Man*, opines, "If a thing is worth winning, it's worth cheating for."[27] As school systems across the country have raised the stakes associated with standardized testing, cheating on these tests has become a tempting option for some teachers and administrators. The investigation for the Chicago Public Schools by Brian Jacob and Steven Levitt has documented cheating by 5 percent or more of the teachers.[28] Their article in this volume describes the system they developed for detecting cheating, based partly on an analysis of patterns of test answers, and provides persuasive validation of that system.

W. C. Fields would not have been surprised by the recent wave of corporate scandals involving accounting manipulations that have the effect of faking profitability, to the great profit of the cheaters. It is also unsurprising that Asian students seeking entrance to American universities would be tempted to fake command of the English language: apparently large numbers of them in recent years have taken advantage of web sites based in China and Korea that posted illegally obtained questions and answers to the verbal part of the Graduate Record Examinations.[29] But it is troubling (to those less cynical than W. C. Fields) that teachers, for whom truthfulness is a professional norm, could succumb to the tempta-

27. Quoted in Kleiman (2002).

28. Jacob and Levitt (2001).

29. Russell Contreras, "Inquiry Uncovers Possible Cheating on GRE in Asia," Associated Press, August (www.irps.ucsd.edu/irps/innews/ap080702).

tion to cheat, especially when the stakes for the individual teachers themselves do not appear that high.

What are the stakes? In the Chicago Public Schools, the superintendent decreed that beginning in 1996 poor performance on the test could result in penalties for both individual students and for entire schools.[30] Students are required to meet minimum standards on the mathematics and reading tests to be promoted from third, sixth, and eighth grades. Schools must have at least 15 percent of their students score above national norms on the reading exam or be placed on probation. Schools placed on probation are threatened with being reconstituted by the central administration if they exhibit insufficient progress in subsequent years. Unlike a number of state systems, Chicago chose not to institute cash bonuses to teachers in schools that did well on the standardized tests.

From the perspective of an individual teacher in Chicago, the incentive to cheat appears dilute. The direct threats are to the individual students (who may be held back) or to entire schools. The performance by any one teacher's students is just a part of the average performance by which her school is judged, and it would be an unusual circumstance that changing some answers for one class would make much difference in the likelihood that her school would end up on probation.[31]

But perhaps that is not the whole story. It would be useful to better understand the incentives to Chicago teachers. One possibility is that teachers *do* have a personal stake in test results, which would be true, for example, if principals would evaluate teachers on the basis of their students' test performance. That would presumably be of greatest concern to new teachers who have not yet received tenure, and suggests that new teachers would be more likely to cheat than established teachers. Another possibility is that teachers cheat out of sympathy for students who would otherwise be held back, which suggests that teachers will be more likely to change the test answers for those students who are likely to fail. These speculations are testable with the Jacob-Levitt data.

Deterrence

Jacob and Levitt have provided the Chicago Public Schools with a tool to help deter future cheating. Their results have triggered an investigation

30. Jacob and Levitt (2001).

31. School administrators have a more direct incentive to encourage or implement cheating, but it appears that most of the "action" in Chicago involved teachers.

by the administration, with penalties likely for several teachers. But if these events are perceived as unique, and unlikely to be repeated, they will have little deterrent value. Is it feasible to construct an ongoing deterrence-based system incorporating the Jacob-Levitt indicators?

To be effective, a deterrence-based system would necessarily include the threat of sanctions for individual teachers. If it is to be accepted by the teachers and their union, it seems essential that it be viewed as fair and reliable. Cost is also a consideration.

Social psychologists thirty years ago demonstrated that the perceived "fairness" of a surveillance system affects the propensity to follow the rules.[32] Perceived fairness may require among other things a degree of transparency. The indicators proposed by Jacob and Levitt, while valid, are difficult to understand and may be hard to sell to the teachers as part of an ongoing system. Even if, as seems likely, the indicators would be used only as the basis for deciding which classrooms required further investigation, that investigation would itself be viewed as punitive by the teachers, and thus would be subject to fairness concerns.

An example of a more transparent indicator is average classroom score on the standardized test relative to expected score (based on past performance of the students). Instances in which a class exceeded expectations by a wide margin would be considered suspect. While Jacob and Levitt have demonstrated that this approach by itself creates more false positives and false negatives than a system that incorporates their indicators, that loss of precision may be a price worth paying.

To economize, investigations could be limited to those classes where the cheating, if it in fact occurred, would likely have made a difference with respect to incurring any of the contingent penalties. That limitation would not affect the deterrent value of this auditing approach.

Prevention

Any deterrence-based system, unless entirely effective, is likely to erode teacher morale. And even if rare, news stories of teachers caught cheating will be costly to the reputation of the school system. Other things equal (including effectiveness and cost), a prevention-oriented system may be preferable.

32. See, for example, Friedland, Thibaut, and Walker (1973).

North Carolina, one of the leaders in high-stakes testing, provides a model. North Carolina public schools are evaluated on the basis of standardized tests, with a more extensive set of contingencies than in the Chicago system. Schools that perform poorly are threatened with intervention by the state. Schools that perform well are rewarded by public recognition, and the teachers in those schools receive cash bonuses of as much as $1,500.[33] The North Carolina Department of Public Instruction anticipated that teachers and school administrators might be tempted to cheat or at least find ways to game the system, so it instituted an extensive set of procedural requirements.[34] The end-of-year test is taken under the supervision of the students' regular teacher, who is observed by a proctor; the proctor must be an adult and is ordinarily not a colleague. Teachers are not allowed to talk to the students during the test, or even to distribute candy or other favors that might serve to improve their mood. Three versions of the test are distributed so that the students cannot copy from each other so easily. Teachers edit the answer sheets for errors in name and other identifiers, but only at a set time and place together with other teachers—otherwise the test sheets are locked up. Teachers cannot have the test booklets with them when editing. The entire school system must take the test on the same day. Attendance must be at least 99 percent (with the denominator carefully defined) for the school to qualify for any honors. And so forth.

Such requirements are a nuisance and might be viewed by some teachers as an insult to their integrity. But they are inexpensive and probably very effective in preventing individual teachers from cheating.[35]

At a more basic level, prevention can be built into the system of contingencies. Helen Ladd observes that "balance must be found so financial awards are large enough to change behavior, but not so large that they induce outright cheating."[36] It is not just the size of the awards that matters, however, but also the contingency system. If cheating is most likely to be a temptation for individual teachers, then perhaps diluting the individual incentive to cheat by tying consequences to the performance of the

33. Ladd and Zelli (2002).

34. These requirements were related to me by Elizabeth Camden Cook, an experienced eighth-grade teacher in the Durham, North Carolina, Public Schools.

35. There may still be room in this system for administrators to cheat, so it does not entirely negate the usefulness of an occasional audit.

36. Ladd (1996, p. 14).

school rather than the classroom would be useful—as has been done in North Carolina, the Chicago Public Schools, and elsewhere. Whether that system will also reduce useful efforts to improve student performance depends on the locus of relevant decisionmaking within the school. To the extent that the relevant features of the instructional process are set by the school or district (for example, choice of texts, pacing, student tracking), then placing the contingency at the school level is entirely appropriate.

Honest Cheating

George Washington Plunkett of Tamany Hall famously distinguished between "dishonest graft" and "honest graft."[37] By today's standard, that might be the distinction between outright bribes as opposed to (legal) influence-buying through campaign contributions. The distinction can usefully be appropriated for high-stakes testing, which can lead to "dishonest cheating" (of the sort documented by Jacob and Levitt) but also "honest cheating." More precisely, "dishonest cheating" means to change the relationship between the performance score and the "true" level of student accomplishment in some way not permitted by the rules. "Honest cheating" means to change the relationship between the score and the true level of accomplishment in some way that is permitted, but defeats the purpose of the system.

"Teaching to the test" is a phrase that suggests some of the most obvious forms of honest cheating. For example, if the purpose of high-stakes testing is to make schools more productive in educating children to read, write, and figure, but the test score is heavily influenced by the students' ability to manage multiple-choice tests effectively, then much instructional time will be devoted to developing test-taking skills at the cost of instruction in substance. Likewise, history, science, physical education, and any other subject that is unmeasured by the test, and therefore outside the ambit of reward and punishment, will be neglected, even if it is generally acknowledged to be important.[38] Furthermore, the type of average or summary score used to rate a school may create an incentive for distributing resources inappropriately; in Chicago, for example, if a school is struggling to avoid probation, the school administration may be

37. See the discussion in Robert D. Behn, "Cheating—Honest & Dishonest," *New Public Innovator* (May–June, 1998), pp. 18–19.

38. Ladd (1996, p. 12).

tempted to focus instructional resources on the most able students to ensure that at least 15 percent reach the national norm.

An ounce or two of prevention may be sufficient to reduce dishonest cheating to some minimal, acceptable level. But there are no cheap remedies available for honest cheating except to do away with high-stakes testing.

Bruce Sacerdote: This is a very exciting and unusual paper that illustrates both how people respond to incentives and how statistical tools can be used to solve real-world problems. The paper is a continuation of Brian Jacob and Steve Levitt's earlier work on the problem of detecting teacher cheating on standardized tests given to secondary school students. In the earlier paper the authors developed several measures of suspicious patterns of answers at the classroom level, and they showed that suspicious patterns were strongly correlated with student test score gains that were not sustained the following year. For the current paper the authors were permitted to choose certain classrooms for auditing and retesting to better ascertain the extent to which suspicious answer patterns were caused by cheating teachers. Below I present several reasons social scientists and policymakers will care about these results and then discuss some implications of the work that the authors did not highlight but might have.

The paper is incredibly interesting for several reasons. First, the use of high-stakes testing is becoming increasingly common and important in public schools. This paper highlights at least one of the potential distortions in behavior from such testing. If students' grade promotion or graduation is tied to their test performance, they will have a strong incentive to cheat, or a teacher might cheat on the students' behalf. Similarly, if jobs and salaries for teachers are tied to aggregate test performance (either at the school or classroom level), the teacher is also given a strong incentive to cheat. Current testing conditions are often far from rigorous. For example, teachers may have access to the answer sheets long after the students have completed the test. And in the case of the Iowa Test of Basic Skills used in the Chicago Public School system, questions are often reused every few years. This practice can allow teachers to give their students some of the actual test questions ahead of time. Jacob and Levitt demonstrate that a number of teachers do take advantage of the lax testing conditions.

Clearly policymakers could respond to this problem by increasing the chance that cheating teachers get caught, and the paper provides a nice set of tools to detect certain forms of cheating. Policymakers could also make it more difficult or time-consuming to cheat with simple steps such as not allowing teachers to proctor their own exams. Cheating of the kind and frequency examined in the paper does not necessarily demonstrate that high-stakes testing is not a viable policy, particularly given that school administrators have a number of low-cost options they could use to reduce cheating. The next interesting direction in this research program will be to see how teacher behavior in Chicago changes following the investigations described in the paper. An interesting question will be whether teachers can substitute more sophisticated cheating strategies that are harder to detect.

One broader message of the paper is that incentives matter and that human beings are inventive in finding ways to game a system. And we should be aware that teachers or police or clergy members are subject to the same economic forces that explain many aspects of human behavior in general. A second broad message is that microdata often contain a great deal of information in the covariance of data items across individual people. In this case the authors show that much can be learned from the extent to which students answers are correlated within a classroom. For example, unusual and large blocks of identical answers from within a classroom may indicate that students cheated from each other or that a teacher filled in portions of the answer sheets. By sifting through the within-classroom correlations of answers, the authors are able to make inferences about student and teacher behavior.

One of my favorite aspects of the paper is the experimental design used for the audit. Rather than have one treatment (suspected cheaters) and one control group, the audit design has three different groups of suspected cheaters and two control groups. Within the suspected cheaters are classrooms with suspicious answer strings and unusual gains in scores, classrooms with suspicious answer strings and normal gains, and classrooms for which there were anonymous accusations of cheating. The retesting of this third group is particularly important for validating the paper's methodology. All three groups of suspected classrooms (including those selected because of anonymous tips that cheating was occurring) experience significant declines in scores on the retest (audit); the control groups do not. Suspected cheaters experience large declines

on the retest even when they are identified through tips rather than by statistical tools. Given these results, one can be fairly confident that the authors' methodology is indeed picking up cheating.

The authors did not discuss what happened in cities that tried to detect cheating through methods other than statistical analysis, but I found reading about such attempts very informative. Gary, Indiana, uncovered a major teacher cheating scandal early in 2002; New York City conducted a similar investigation in 1999. The investigation in New York City was started when a high school that was in danger of being closed because of poor test scores suddenly experienced large score gains. Investigators interviewed a host of witnesses, including students and teacher's aides, and found actual cheat sheets that students had been handed. These instances demonstrate that statistical sophistication is not a precondition to being able to catch cheaters.

However, some of the New York classrooms accused of cheating retained their large improvements in scores, and it is possible that some of the teachers who were reassigned or dismissed were simply those who had succeeded in improving their students' test-taking ability. This possibility illustrates the danger and irony of using score improvements alone to determine who gets investigated, and it speaks further to the value of the current paper.

Do cheaters prosper? Based on the retest, students in cheating classrooms still experience average test score increases, even after one removes their ill-gotten gains from cheating. In fact, on the mathematics retest, students in cheating classrooms had better annual gains than the average student in the Chicago public schools. On the initial test in 2002 the most likely cheaters experienced an annual gain of 30.0 points on the mathematics test versus the systemwide average gain of 16.9 points. On the retest the most likely cheaters still racked up a gain of 19.3 points, though this gain is probably not statistically significantly greater than the 16.9 average gain. This may indicate that the most aggressively cheating teachers are not necessarily short-changing their students in the classroom. In fact, if the teacher is highly motivated, she might use multiple ways to try to increase scores, including both approved and disapproved methods.

Overall the paper makes large contributions on various levels. First, it gives school administrators and researchers tools for detecting cheating. Second, it shows one of the potential pitfalls of tying teachers' jobs or

students' promotions to standardized tests. With regard to this point, the evidence suggests that schools are currently doing little to prevent teachers from cheating. Therefore low-cost increases in deterrence may cause large reductions in the frequency and extent of cheating. Third, the paper is an inspiration to other researchers who want to do policy-oriented work that has some near-term impact on the world. The paper shows that collaborative efforts between researchers and practitioners can be highly productive. Finally, it reminds us that people respond to incentives in very rational ways and that social science can tell us a great deal about some aspects of human behavior.

References

Angoff, William H. 1974. "The Development of Statistical Indices for Detecting Cheaters." *Journal of the American Statistical Association* 69 (345): 44–49.

Cizek, Gregory J. 1999. *Cheating on Tests: How to Do It, Detect It and Prevent It.* Lawrence Erlbaum Associates.

Deere, Donald, and Wayne Strayer. 2001. "Putting Schools to the Test: School Accountability, Incentives and Behavior." Working Paper 0113. Department of Economics, Texas A&M University.

Frary, Robert B., T. Nicolaus Tideman, and Thomas Morton Watts. 1977. "Indices of Cheating on Multiple-Choice Tests." *Journal of Educational Statistics* 2 (4): 235–56.

Friedland, Nehemia, John Thibaut, and W. Laurens Walker. 1973. "Some Determinants of the Violation of Rules." *Journal of Applied Social Psychology* 3 (2): 103–18.

Grissmer, David W., and others. 2000. *Improving Student Achievement: What NAEP Test Scores Tell Us.* MR-924-EDU. Santa Monica: RAND Corporation.

Heubert, Jay P., and Robert M. Hauser, eds. 1999. *High Stakes: Testing for Tracking, Promotion and Graduation*. National Academy Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design." *Journal of Law, Economics and Organization* 7 (Spring): 24-51.

Jacob, Brian A. 2001. "Getting Tough? The Impact of Mandatory High School Graduation Exams on Student Outcomes." *Educational Evaluation and Policy Analysis* 23 (2): 99–121.

———. 2002. "The Impact of Test-Based Accountability in Schools: Evidence from Chicago." Unpublished manuscript. John F. Kennedy School of Government, Harvard University.

Jacob, Brian A., and Steven D. Levitt. Forthcoming. "Rotten Apples: An Estimation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics.*

Kleiman, Mark A. R. 2002. "Dukenfield's Law of Incentive Management." Department of Policy Studies, University of California at Los Angeles.

Klein, Stephen P., and others. 2000. "What Do Test Scores in Texas Tell Us?" Santa Monica: RAND.

Ladd, Helen F., ed. 1996. "Introduction." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 1–21. Brookings.

Ladd, Helen F., and Arnaldo Zelli. 2002. "School Based Accountability in North Carolina: The Responses of School Principals." *Educational Administration Quarterly* 38 (October): 494–529.

Loughran, Regina, and Thomas Comiskey. 1999. "Cheating the Children: Educator Misconduct on Standardized Tests." Report of the City of New York Spe-

cial Commisioner of Investigation for the New York City School District (December).

Richards, Craig E., and Tian Ming Sheu. 1992. "The South Carolina School Incentive Reward Program: A Policy Analysis." *Economics of Education Review* 11 (1): 71–86.

Smith, Stephen S., and Roslyn A. Mickelson. 2000. "All that Glitters Is Not Gold: School Reform in Charlotte-Mecklenburg." *Educational Evaluation and Policy Analysis* 22 (2): 101–27.

Tepper, Robin Leslie. 2001. *The Influence of High-Stakes Testing on Instructional Practice in Chicago*. Seattle: American Educational Research Association.

Van der Linden, Wim, and Leonardo Sotaridona. 2002. "A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests." University of Twente, Netherlands.