NBER WORKING PAPER SERIES

CATCHING CHEATING TEACHERS: THE RESULTS OF AN UNUSUAL EXPERIMENT IN IMPLEMENTING THEORY

Brian A. Jacob
Steven D. Levitt

Working Paper 9414
http://www.nber.org/papers/w9414

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

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, CA. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory
Brian A. Jacob and Steven D. Levitt
NBER Working Paper No. 9414
December 2002
JEL No. I20, K42

ABSTRACT

This paper reports on the results of a prospective implementation of methods for detecting teacher cheating. In Spring 2002, over 100 Chicago Public Schools elementary classrooms were selected for retesting based on the cheating detection algorithm. Classrooms prospectively identified as likely cheaters experienced large test score declines. In contrast, classes that had large test score gains on the original test, but were prospectively identified as being unlikely to have cheated, maintained their original gains. Randomly selected classrooms also maintained their gains. The cheating detection tools were thus demonstrated to be effective in distinguishing between classrooms that achieved large test-score gains as a consequence of cheating versus those whose gains were the result of outstanding teaching. In addition, the data generated by the implementation experiment highlight numerous ways in which the original cheating detection methods can be improved in the future.

Brian Jacob
Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
brian_jacob@harvard,edu

Steven Levitt
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
slevitt@midway.uchicago.edu

Most large urban school districts in the United States suffer from low test scores, high dropout rates, and frequent teacher turnover. In response to these concerns, the last 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 (Jacob 2002, Grissmer et. al. 2000), critics have argued that these gains are artificially induced by "teaching to the test." 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 (Jacob 2002, Klein et. al. 2000). Even more ominous is the possibility that the emphasis on high-stakes testing induces cheating on the part of students, teachers, and administrators.

Jacob and Levitt (2002) develop a method for detecting cheating by teachers and administrators on standardized tests. The basic idea underlying that method (which is described in greater detail in Section II) 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, however, as an indicator of cheating, are tell-tale patterns of suspicious answer strings, e.g. 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. Jacob and Levitt (2002) conclude that cheating occurs in 3-5 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 of this paper. At the invitation of Arne Duncan, CEO of the Chicago Public Schools, we were granted the opportunity to work with CPS administration to design and implement auditing and retesting procedures implementing the tools developed in Jacob and Levitt (2002). Using that 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 third to eighth grade. The classrooms retested include not only cases suspected of cheating, but also classrooms 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 developed by Jacob and Levitt.

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. In contrast, classrooms not suspected of cheating a priori 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 cheating detection techniques. The cheating algorithm was developed without access to multiple observations for the same classrooms. By observing two sets of results from the same classroom (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 others. By changing the weights used in the algorithm, we should be able to substantially improve the predictive value of the model in future implementations.

The remainder of the paper is structured as follows. Section II presents background information on teacher cheating and the detection methods. Section III outlines the design and implementation of the retesting procedure. Section IV reports the results of the retests. Section V uses the data from the retests to analyze the predictive value of the various components of the algorithm and identifies a number of improvements to the methods. Section VI concludes.

Section II: Background on 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. Prior to the passage of that law, every state in the country except for Iowa already administered state-wide assessment tests to students in elementary and secondary school. Twenty-four states require students to pass an exit examination to graduate high school. In the state of California, a policy providing for merit pay bonuses of as much as $25,000 per teacher in schools with large test score gains was recently put into place.

Critics of high-stakes testing argue that linking incentives to performance on standardized tests will lead teachers to substitute away from other teaching skills or topics not directly tested on the accountability exam (Holmstrom and Milgrom 1991). 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 non-tested areas.2

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.3 While the idea of elementary school teachers manipulating student answer sheets may seem far-fetched, cheating scandals have been appeared in many places including California (May 2000), Massachusetts (Marcus 2000), New York (Loughran and Comiskey 1999), Texas (Kolker 1999), and Great Britain (Tysome 1994). Jacob and Levitt (2002) provide the first systematic analysis of teacher cheating.4 We argue in that paper that cheating classrooms are likely to share three characteristics: (1) unusually large test score gains for students in the class the year the cheating occurs, (2) unusually small gains the following year for those same students, and (3) distinctive patterns of "suspicious" answer strings.

The first two characteristics above relating to test scores are straightforward. Large increases are expected in cheating classrooms because raising test scores is the very reason for the

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

3. 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.

4. In contrast, there is a well-developed literature analyzing student cheating (e.g., Aiken 1991, Angoff 1974, Frary, Tideman, and Watts 1977, van der Linden 2002).

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 the following year.

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 to be 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 non-consecutive questions in order 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 utilize for separate suspicious string measures to try to detect all of these varieties of cheating. 5 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 socio-economic characteristics as explanatory variables. Past test scores, particularly on the same subject test, are very powerful predictors of the student answers on the current test.

The first suspicious string indicator used is a measure of how likely it is that, by chance,

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

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 maps directly into the most naive form of cheating highlighted above, but may not adequately identify more sophisticated types of cheating, which are addressed by our second and third measures. The second indicator measures the overall degree 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 has a few questions in which the correlation in student answers are quite high, but the degree of correlation across students in the classroom on other questions is unremarkable, this potentially suggests intervention on the part of the teacher on the questions in which answers are highly correlated. The fourth and final suspicious string indicator 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 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.

In order to construct an overall summary statistic measuring the degree of suspiciousness of a classroom's answers, we rank order the classes from least to most suspicious within subject and grade on each of the four individual 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 (i.e., the most suspicious part) of the distribution. A parallel statistic is constructed for the two test-score gain measures corresponding to this 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, Jacob and Levitt (2002) present a wide range of evidence supporting the argument that these measures have predictive power empirically. For instance, among classrooms that have large test score gains this year, children in classrooms that have of 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 to be due to cheating. Also, there tends to be strong correlations across subjects within a classroom, 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 the set of incentives that are in place. For example, cheating is more likely to occur in low-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. Figure 1 presents a graph in which 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 large test score gain in the current year followed by an unexpectedly small increase (or even a decline) in the following year.6 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

6. 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.

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 reflects the fact that teacher cheating increases the likelihood of suspicious strings and of large test score jumps. In Jacob and Levitt (2002), we formally demonstrate that under a set of carefully articulated assumptions, the area under the curve in Figure 1 above the projection one would make based on observing the left-hand portion of the figure captures the overall incidence of teacher cheating. Empirically, our findings imply that as many as 5 percent of the classrooms in CPS show evidence of cheating on the ITBS in any given year.

-----Figure 1 about here-----

Section III: Implementation of the cheating detection algorithm in the Spring 2002 ITBS testing

Each Spring, roughly 100,000 CPS students take the ITBS test. The results of this test determine (1) which schools will be placed on academic probation or reconstituted, (2) which students will be required to attend summer school and potentially be retained ($3^{rd,}$ 6^{th}, and 8th grade only), and (3) what students are eligible to apply to the most sought after test-based magnet high schools in the CPS system (7^{th} grade).

The accountability department CPS conducts retests of the ITBS in roughly 100 classrooms annually for the purpose of quality assurance. The retests, which use a different version of the exam, occur 3-4 weeks after the initial testing. Specially trained staff in the accountability office administers the retests. Unlike the initial round of testing, which is subject to relatively lax oversight and control and potentially affords a variety of school staff access to the test booklets,

the retest answer sheets are closely guarded. Up until the last few years, classrooms were randomly selected for retests.7 In recent years, retests have been focused on those classrooms achieving the largest test score gains relative to the prior year. Formal investigations have been undertaken when major discrepancies arise between the official testing and the retest, but punishment is extremely 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 implementation of the audits was that budget limitations restricted the total number of classrooms audited to be no more than 120. It is important to note that our earlier research on cheating estimated that there were roughly 200 classrooms cheat each year in CPS. Thus, the budget constraint meant that we were able to audit only a fraction of suspected cheaters.

Selecting individual classrooms with the goal of *prospectively* identifying cheating raised an important issue since our original cheating detection method developed in Jacob and Levitt (2002) relies heavily on 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). In selecting classrooms to retest, however, next year's test scores did not yet 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 this year.

---

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

Table 1 lays out the structure of the implementation scheme that was developed. Classrooms to be audited were divided into five separate categories. The first set of classrooms exhibited both unusually large test score gains and highly suspicious patterns of answer strings. These classrooms were *a priori* judged to be the most likely to have experience cheating. A second group of classrooms had very suspicious answer string patterns, but did not have unusually large test score gains. That pattern is consistent with a bad teacher who failed to adequately teach the students and attempted to cover up this fact by cheating. Thus, these classrooms were *a priori* suspected of high rates of cheating. A third set of classrooms were those for which anonymous allegations of cheating were made to CPS officials. There were only four such classrooms. It is worth noting that none of these four classes accused of cheating would have otherwise made the cutoff for inclusion in our first two groups of suspected cheaters. The remaining two types of classrooms audited were not suspected of cheating, but rather, served as control groups. One set of controls were classrooms with large test score gains, but answer string patterns that did not point to cheating. These classrooms were judged as likely to have good teachers, capable of generating big test score gains without resorting to devious means.8 As such, they provide an important comparison group to the suspected cheaters with large gains. A fifth and final set of classrooms were randomly chosen from all remaining classrooms. These classrooms are also unlikely to have high rates of cheating.

-----*Table 1 about here* -----

With the exception of the anonymous tips and the randomly chosen group, we did not

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

employ a hard and fast cutoff rule for allocating classrooms into the various categories. In order to be assigned to the first or second category, a classroom generally needed to be in the top few percent of classrooms on suspicious answer strings on at least one subject test. For category 1, 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 suspiciousness, these cutoffs were sometimes relaxed. In addition, some classrooms that appeared suspicious, but otherwise would not have made it into categories 1 or 2, were included because other classrooms in the same school did qualify and we were interested in isolating school-wide 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 with large achievement gains but that 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 suspected cheating classrooms is due to reduced effort on the part of students on the retest.9 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 among the high-achieving, suspected cheaters may simply be the consequence of mean reversion. Second, including the control groups allows us to more effectively test how various components of our model are working in identifying cheating after the fact. The cost of the retest structure we implemented is that the inclusion of control groups means that we are able to retest fewer classrooms suspected of cheating. Of the 117 classrooms retested,

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

76 were suspected of cheating (51 with suspicious strings and large test score gains, 21 with only suspicious strings, and 4 anonymous tips). As noted above, there were many more classrooms that looked equally or nearly as suspicious, but were not retested due to resource constraints.10

In some cases, classrooms were retested on only the math or the reading subject tests, not both.11 In particular, classrooms that were suspected of cheating only on math 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 math or the reading test was administered, but never both. In the results presented below, we only report test score comparisons for those subjects on which retests took place.

Section IV: Results of the Retests

The basic results of the retests are presented in Table 3. For most of the categories of classrooms defined above, six different average test score gains are presented (three each for math and reading).12 For the randomly selected classrooms, there is so little data that we lump together math and reading. For the classes identified by anonymous tips, audits took place only on reading

10. 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.

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

12. Whenever we talk about test score gains, we are referring to the change in test scores for a given student, on tests taken at different points in time.

so we do not report math scores. In all cases, we report the test score gains in terms of standard score units, the preferred metric of the CPS. A typical student gains approximately 15 standard score units per academic year.

-----*Table 2 about here*-----

In columns 1-3, we report the results on the reading subject test (and the combined reading and math test results for the randomly selected classrooms). Column 1 presents test scores between Spring 2001 and the Spring 2002 ITBS (the actual test, not the retest). Among all classrooms in CPS (both those that are retested and those that are not), the average gain on the reading test was 14.3 standard score points. Classrooms *a priori* identified 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 the previous year. Our control group of good teachers achieved gains that were large (20.6), but not as great as the suspected cheaters. Bad teachers suspected of cheating had test score gains slightly above the average CPS classroom. The randomly selected classes were in line with the overall CPS, 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 also saw large declines of 8.8 standard score points. The anonymous tip classes lost 6.8 points. In stark contrast, however, the good teacher classrooms actually register small *increases* on the audit test relative to the original.13 The randomly selected classrooms lost 2.3 points, or only one-seventh as much as the most likely cheaters. The fact that

13. As noted above, math 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.

the two control groups (good teachers and randomly selected classes) saw only small declines suggests that any impact of decreased effort by students on the retest are 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 due to cheating.

Column 3 reports 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 sum of column 1 and 2).14 The largest "true" gains, as would be expected, are in the classrooms identified as good teachers. The most likely cheater classes that scored so high on the initial test, look merely average in terms of "true" gains, suggesting 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 of 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 4-6 report results parallel to the first three columns, but for math instead of reading. The results are generally similar to those for reading, but less stark.15 The good teachers

14. Subject to the caveat that effort might have been lower on the retest and that the Spring 2001 scores might themselves be inflated by cheating that occurred in the prior year.

15. A partial explanation for why the results on the math test are less stark than those for reading is that the math test is made up of three separate parts, unlike reading, which is in one self-contained section. In conducting the retests, classrooms suspected of cheating on any of the three math sections were retested on the entire math test. Thus, included in the math results are some classes where there was strong evidence of cheating on one part of the math exam, but not on

have baseline math gains commensurate with 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 (over 10 standard score point declines in each case). The good teacher classrooms also see a small decline in math scores on the retest (3.3 standard score points), unlike on reading where they gained. Finally, in column 6, a notable difference between the results for reading and math is that the classrooms *a priori* judged most likely to be cheating showed above average "true" gains on math, which was not the case for reading. This result is likely due to the fact that our modified algorithm used for *prospectively* identifying cheaters relies in part on large test score gains, and thus is biased towards 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.16

Figures 2 and 3 present the cumulative distribution of changes in test scores between the initial Spring 2002 test and the retest for classrooms in different categories for reading and math

---

another part. Even when the math results are further disaggregated, identifying particular sections of the math exam where classes were *a priori* judged likely to have cheated, the results are not as clean as for reading.

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

respectively. These figures highlight the stark differences between the classes *a priori* predicted to be cheating and those identified as good teachers. The vertical axis is the cumulative percent of classrooms with a test score change between the initial test and the audit that is less than the value named on the horizontal axis. Three separate cumulative distributions are plotted in each figure corresponding to the *a priori* most suspicious classrooms, bad teachers suspected of cheating, and good teachers. The striking feature of the figures is how little overlap there is between the cheating and good teacher distributions. 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 did not have a long left tail like the most suspicious cheaters, but had a high concentration of cases in which there were double-digit losses. In contrast, the single biggest test score decline experienced by a good-teacher classroom on reading is seven standard points (as indicated by the cumulative distribution rising above zero at that point for the good teacher curve). More than eighty percent of the most suspicious classrooms experienced losses greater than that, and almost 60 percent of bad teacher classrooms saw bigger declines. Note also that about one-third of the good teacher classrooms experienced test score gains, whereas virtually none of the suspected cheating classrooms did.

-----Figure 2 about here-----

The results in Figure 3 are similar. The primary differences between the two figures are that (1) the distribution of outcomes for the most suspicious teachers and the bad teachers suspected of cheating are almost identical on the math test, and (2) 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 potentially identifying cheating classrooms, but also, they are at least as successful in identifying classrooms with good teachers whose gains are legitimate and are possibly deserving of rewards, as well as focused analysis as examples of best practices.17

-----Figure 3 about here -----

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?18 Unfortunately, the way in which the audits were implemented limit the amount of light we are able to shed on this issue. The CPS officials who determined which classrooms to audit intentionally tried to avoid retesting large numbers of classes in individual schools due to the negative reaction that elicits in the schools. There are at least two schools, however, in which the

17. Although some caution must be exercised in discussing "good" teachers. Are findings suggest that classrooms with big test score gains that do not have suspicious answer string patterns can maintain their gains on retests. Whether the large test score gains are the result of artificially low test scores in the prior year (due perhaps to a previous bad teacher or adverse test conditions in the preceding year) is not something we have explored.

18. 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/culture that encourages cheating, or systematic differences in incentives across schools (e.g. because low performing schools are threatened with probation and reconstitution).

audits provide 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 the bulk of the cheating incidents are consistent with teachers rather than administrators doing the cheating.

Section V: Using the retests to evaluate and improve cheating detection

Up to this point, the paper has focused on evaluating how effective the methods previously developed were in prospectively identifying 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 different indicators discussed earlier. Each of the four indicators is 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:

$$Change\_in\_test\_score_{cs} = Suspicious\_string\_measures_{cs}'\Gamma + \theta_s + \theta_g \quad (1)$$

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 of interest are the suspicious string measures, which will be entered in a variety of different 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 different subject tests (reading comprehension and three math 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 equation 1) is a highly selected 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 we use two indicator variables corresponding to whether a classrooms is in the 99^{th} percentile on this measure, or between the 90^{th} and 99^{th} percentiles. We have experimented with a fuller parameterization, but this sparse specification appears to adequately capture the relevant variation. Classrooms in the 99^{th} 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 90^{th} percentile). This result is highly statistically significant. Classes in the 90^{th}-98^{th} percentiles lose only one-third as much, although

the result is still statistically significant.[19] Thus, there appears to be a sharp discontinuity occurring in the last one 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.

-----*Table 3 about here*-----

Column 2 adopts a different functional form for the suspicious answer strings measure. Rather than aggregating over the four different 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 very extreme on particular measures (although possibly not extreme at all on other ones) relative to classrooms that are somewhat elevated on all four measures. Being in the 99th percentile on all four measures individually – a very 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. While there is a large jump between being in the 99th percentile on all four measures versus 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. Note that 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.

19. If one allows the impact of 90th-94th percentile to differ from 95th-98th, one cannot reject that the coefficients are identical on those two variables. Indeed, the point estimate on 90th-94th is slightly larger than that on 95th-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 than zero. In contrast, the indicator variables for the separate measures continue to enter strongly and with a similar pattern as before. The R-squared of the nested model in colum 3 is only slightly above 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 are 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 an 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 is the information contained in the answer strings to the actual subject test.

Columns 5-8 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.20

20. We are guarded in our interpretation of this coefficient and these specifications in general,

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-98th percentiles. 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 statistically significant, and one cannot reject equality of coefficients between being in the 99th percentile and the 90th-98th percentile. A high variance in the degree 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.

-----*Table 4 about here*-----

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

---

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 (i.e. good teachers and randomly selected controls).

on the measure reflecting students getting the hard questions right and the easy ones wrong, and by de-emphasizing or eliminating altogether the measure of variance across questions.

Section VI: Conclusions

This paper summarizes the results of a unique policy implementation that allowed a prospective test of cheating detection tools we had previously developed. The results of retests generally support the validity of these tools for prospectively identifying teacher cheating. Classrooms *a priori* selected as likely cheaters saw dramatic declines in scores on retests, whereas classes identified as good teachers and 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 it even more effective in future applications.

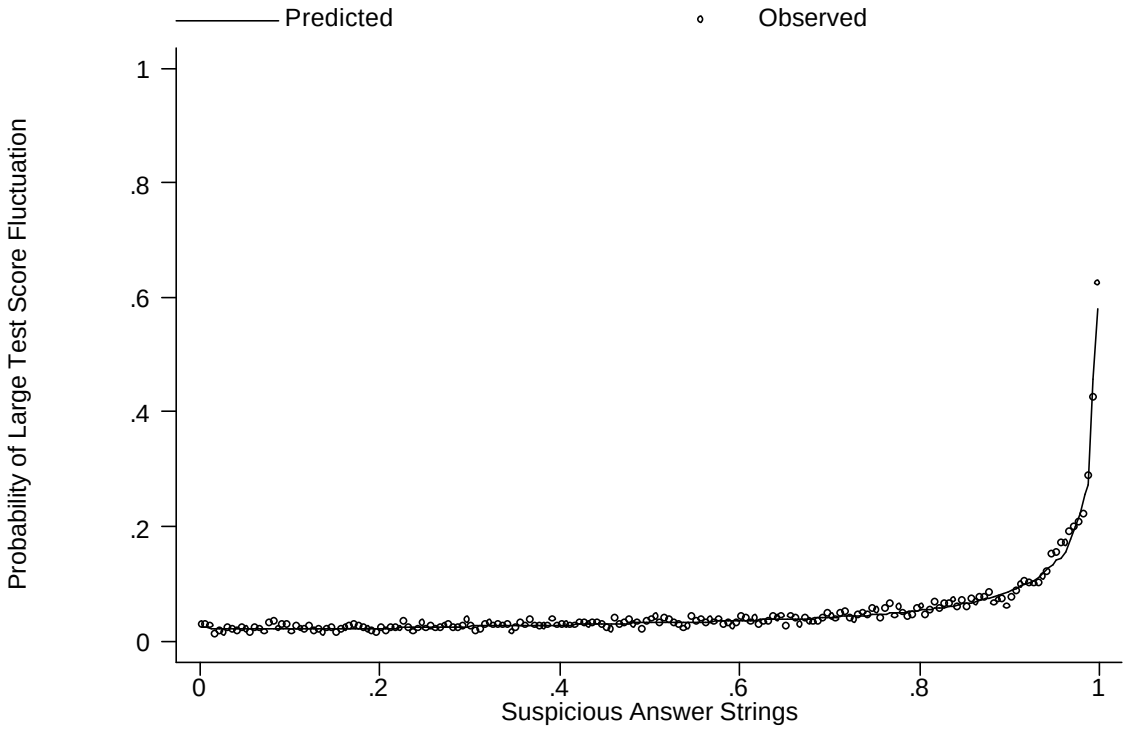
On a more practical level, the implementation demonstrated the value of these tools to school districts interested in catching cheaters or deterring future cheating. Out of almost 7,000 potential classrooms, our methods isolated 70 suspicious classrooms that were retested (as well as many more equally suspicious classrooms that were not retested due to budget constraints). Of these seventy, almost all experienced substantial declines on the retest indicative of cheating. In 29 classrooms, the test score declines were particularly great (more than one grade-equivalent on average across the subjects retested). CPS staff further undertook further investigation of these 29 classrooms, including analysis of erasure patterns and on-sight investigations. Although the outcome of disciplinary actions is 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 towards 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 1-2 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. On the other hand, 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 tradeoff between the real benefits of high-stakes testing and the real costs associated with behavioral distortions aimed at artificially gaming the standard.

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 zero representing the least suspicious classroom and one representing the most suspicious classroom. The 95th percentile 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.

Table I: Design of the 2002 Sample of Classrooms to be Audited

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 phoned in 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	A control group	NO	NO	Little change between original test and audit	24

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

Table II: Results of Retesting: Comparison of Results for Spring 2002 ITBS and Audit Test

Category of classroom	Reading gains between...			Math gains between...		
	Spring 2001 and Spring 2002	Spring 2002 and 2002 retest	Spring 2001 and 2002 retest	Spring 2001 and 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

Notes: Because of limited data, math and reading results for the randomly selected classrooms are combined. Only the first two columns are available for all CPS classrooms since audits were performed only on a subset of classrooms. All entries in the table are in standard score units.

Table III: The Relationship between Suspicious Answer Strings and Score Declines on the Retest

Measure of suspicious answer strings	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Overall measures of suspicious answer strings (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)
R-squared	.462	.518	.530		.512	.559	.569	.582

Notes: 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 set of 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 clustered to take into account correlation within classrooms across different subject tests.

Table IV: The Performance of the Individual Suspicious String Indicators
in Predicting Score Declines on the Retest

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)
R-squared	.482	.524	.529	.558

Notes: 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 set of 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 clustered to take into account correlation within classrooms across different subject tests.

References

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

Cizek, G. J. (1999). *Cheating on Tests: How to Do It, Detect It and Prevent It.* New Jersey: Lawrence Erlbaum Associates.

Deere, D. and W. Strayer. 2001. "Putting Schools to the Test: School Accountability, Incentives and Behavior." Working paper. Department of Economics, Texas A&M University.

Frary, R.B., Tideman, T.N. and Watts, T.M. 1977. "Indices of cheating on multiple-choice tests." *Journal of Educational Statistics*, 2, 235-256.

Heubert, J. P. and R. M. Hauser, Eds. 1999. *High Stakes: Testing for Tracking, Promotion and Graduation*. Washington, D.C., National Academy Press.

Grissmer, D.W. et. al. 2000. Improving Student Achievement: What NAEP Test Scores Tell Us. MR-924-EDU. Santa Monica: RAND Corporation.

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.

Jacob, Brian A. 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. 2001. Rotten Apples: An Estimation of the Prevalence and Predictors of Teacher Cheating." Mimeo, University of Chicago Department of Economics..

Klein, S. P., L. S. Hamilton, et al. 2000. "What Do Test Scores in Texas Tell Us?" Santa Monica, CA: RAND.

Kolker, Claudia. 1999. "Texas Offers Hard Lessons on School Accountability." *Los Angeles Times*, April 14, 1999.

Loughran, Regina, and Thomas Comiskey. 1999. "Cheating the Children: Educator Misconduct on Standardized Tests." Report of the City of New York Special Commisioner of Investigation for the New York City School District, December.

Marcus, John. 2000. "Faking the Grade." *Boston Magazine*, February.

May, Meredith. 1999. "State Fears Cheating by Teachers." *San Francisco Chronicle*, October 4.

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

Smith, S. S. and R. A. Mickelson. 2000. "All that Glitters is Not Gold: School Reform in Charlotte-Mecklenburg." *Educational Evaluation and Policy Analysis* 22(2): xxx.

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

Tysome, T. 1994. Cheating purge: Inspectors out. *Times Higher Education Supplement*, p. 1, August 19.

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