NBER WORKING PAPER SERIES

ROTTEN APPLES: AN INVESTIGATION OF THE PREVALENCE AND PREDICTORS OF TEACHER CHEATING

Brian A. Jacob
Steven D. Levitt

Working Paper 9413
http://www.nber.org/papers/w9413

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

We would like to thank Suzanne Cooper, Mark Duggan, Sue Dynarski, Arne Duncan, Michael Greenstone, James Heckman, Lars Lefgren, and seminar participants too numerous to mention for helpful comments and discussions. We also thank Arne Duncan, Phil Hansen, Carol Perlman, and Jessie Qualles of the Chicago Public Schools for their help and cooperation on the project. Financial support was provided by the National Science Foundation and the Sloan Foundation. All remaining errors are our own. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating

Brian A. Jacob and Steven D. Levitt

NBER Working Paper No. 9413
December 2002
JEL No. I20, K42

ABSTRACT

We develop an algorithm for detecting teacher cheating that combines information on unexpected test score fluctuations and suspicious patterns of answers for students in a classroom. Using data from the Chicago Public Schools, we estimate that serious cases of teacher or administrator cheating on standardized tests occur in a minimum of 4-5 percent of elementary school classrooms annually. Moreover, the observed frequency of cheating appears to respond strongly to relatively minor changes in incentives. Our results highlight the fact that incentive systems, especially those with bright line rules, often induce behavioral distortions such as cheating. Statistical analysis, however, may provide a means of detecting illicit acts, despite the best attempts of perpetrators to keep them clandestine.

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

I. Introduction

High-stakes testing has become an increasingly prominent feature of the educational landscape. Every state in the country except for Iowa currently administers 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. Twenty states reward schools on the basis of exemplary or improved student performance on standardized exams and 32 states sanction schools on the basis of poor student performance on these exams. 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. Recent federal legislation promises to accelerate this trend. The reauthorization of the Elementary and Secondary Education Act (ESEA) requires states to test students in third through eighth grade each year and to judge the performance of schools based on student achievement scores.

Proponents of high-stakes testing argue that requiring students to demonstrate proficiency in basic skills provides increased incentives for learning, as well as preventing unqualified students from being promoted to higher-level grades where their inadequate preparation may interfere with other students' learning. By linking teacher salary and employment to student test scores, schools are held accountable for their students' performance. Opponents of test-based accountability, on the other hand, argue that linking incentives to performance on standardized tests may unfairly penalize certain students and 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.[1]

In this paper, we explore a very different concern regarding high-stakes testing—cheating on the part of teachers and administrators.[2] As incentives for high test scores increase, unscrupulous teachers may be more likely to engage in a range of illicit activities, including changing student responses on answer sheets, filling in the blanks when a student fails to complete a section, allowing students extra time to complete tests, providing correct answers to students, or obtaining copies of an exam illegitimately prior to the test date and teaching students using knowledge of the precise exam questions. While such allegations may seem far-fetched, documented cases of such cheating have recently been uncovered in California (May 2000), Massachusetts (Marcus 2000), New York (Loughran and Comiskey 1999), Texas (Kolker 1999), and Great Britain (Hofkins 1995, Tysome 1994).

Although the absolute number of teachers and administrators who have been caught cheating to date is very small, there are indications that the prevalence of cheating may be far more widespread. A survey of elementary school teachers in two large school districts asked teachers to what extent they believed an array of questionable actions were practiced by teachers in their school. Almost ten percent of the teachers responded that they believed that teachers in their school "often" or "frequently" give students answers to test questions. Six percent of the respondents believed that teachers "often" or "frequently" changed answers on a student's answer sheet (Shephard and Doughtery 1991). In another study, 35 percent of North Carolina

[1] See, for example, Deere and Strayer (2001), Grissmer et. al. (2000), Heubert and Hauser (1999), Jacob (2001a, 2001b), Klein et. al. (2000) Richards and Sheu (1992), Smith and Mickelson (2000), and Tepper (2001).

[2] Hereafter, we uses the phrase "teacher cheating" to encompass cheating done by either teachers or administrators.

teachers in grades 3, 6, 8 and 10 reported having witnessed cheating, including giving extra time on tests, changing students' answers, suggesting answers to students and directly teaching sections of the test (Gay 1990).

Nonetheless, there has been very little previous empirical analysis of teacher cheating.[3] The few studies that do exist involve investigations of specific instances of cheating and generally rely on the analysis of erasure patterns and the controlled re-testing of students.[4] While this earlier research provides convincing evidence of isolated cheating incidents, our paper represents the first systematic attempt to (1) identify the overall prevalence of teacher cheating empirically and (2) analyze the factors that predict cheating. To address these questions, we use detailed administrative data from the Chicago Public Schools (CPS). In particular, for the years 1993-2000, we have the question-by-question answers given by every student in grades 3-7 taking the Iowa Test of Basic Skills (ITBS).[5] This test is administered annually to virtually all elementary school students in the CPS. In addition to the test responses, we also have access to

[3] In contrast, there is a well-developed statistics literature for identifying whether one student has copied answers from another student (Wollack 1997; Holland 1996; Frary 1993; Bellezza and Bellezza 1989; Fray, Tideman and Watts 1977; Angoff 1974). These methods involve the identification of unusual patterns of agreement in student responses and, for the most part, are only effective in identifying the most egregious cases of copying. Educational Testing Services (ETS), the company that administers national tests such as the SAT, LSAT, and GRE, has funded much of this research (Cizek 1999).

[4]. In the mid-eighties, Perlman (1985) investigated suspected cheating in a number of Chicago public schools (CPS). The study included 23 suspect schools— identified on the basis of a high percentage of erasures, unusual patterns of score increases, unnecessarily large orders of blank answer sheets for the ITBS and tips to the CPS Office of Research—along with 17 comparison schools. When a second form of the test was administered to the 40 schools under more controlled conditions, the suspect schools did much worse than the comparison schools. An analysis of several dozen Los Angeles schools where the percentage of erasures and changed answers were unusually high revealed evidence of teacher cheating (Aiken 1991). One of the most highly publicized cheating scandals involved Stratfield elementary, an award-winning school in Connecticut. In 1996, the firm that developed and scored the exam found that the rate of erasures at Stratfield was up to five times greater than other schools in the same district and that 89 percent of erasures at Stratfield were from an incorrect to a correct response. Subsequent re-testing resulted in significantly lower scores (Lindsay 1996).

[5] We do not, however, have access to the actual test forms that students filled out so we are unable to analyze these tests for evidence of suspicious patterns of erasures.

each student's full academic record, including past test scores, the school and room to which a student was assigned, special education status, free-lunch eligibility, race, gender, and age.

Our approach to detecting classroom cheating uses two types of indicators: unexpected test score fluctuations and unusual patterns of answers for students within a classroom. Teacher cheating increases the likelihood that students in a classroom will experience large, unexpected increases in test scores one year, followed by very small test score gains (or even declines) the following year. Teacher cheating, especially if done in an unsophisticated manner, is also likely to leave tell-tale signs in the form of blocks of identical answers, unusual patterns of correlations across student answers within the classroom, or unusual response patterns within a student's exam (e.g., a student who answers a number of very difficult questions correctly while missing many simple questions).

Empirically, however, not every classroom with test score fluctuations and suspicious answer strings is cheating. Sometimes such patterns arise by chance. To identify the number of cheating classrooms, we would like to compare the observed distribution of test score fluctuations and suspicious answer strings to a counterfactual in which no cheating occurs. Because we do not have the luxury of observing this counterfactual, we must instead make assumptions about what the patterns would look like absent cheating. Our identification strategy hinges on three key assumptions: (1) cheating increases the likelihood a class will have large test score fluctuations and suspicious answer strings, (2) if cheating classrooms had not cheated, their distribution of test score fluctuations and answer strings would be identical to non-cheating classrooms, and (3) the same pattern of correlation between test score fluctuations and suspicious answers observed for non-cheating classrooms in other parts of the distribution also holds in the

upper tail of the distribution. If these assumptions hold, then we can use part of the observed distribution of outcomes that is unlikely to contain many cheaters (e.g. the 50th-75th percentile of suspicious answer strings) to determine the natural correlation between test score fluctuations and suspicious answer strings in non-cheating classrooms. That allows us to predict the patterns we would expect to observe in the tail of the distribution if no cheating occurred. The gap between the predicted and observed frequency of classrooms that are extreme on both the test score fluctuation and suspicious answer string measures provides our estimate of cheating. Because this identification strategy is necessarily indirect, we devote a great deal of space in the paper to presenting a wide variety of tests attempting to confirm the validity of our approach, the sensitivity of the results to alternative assumptions, and the plausibility of our findings.

Figure I provides a simple visual means of demonstrating the empirical approach taken. The horizontal axis in the figure ranks classrooms according to how suspicious their answer strings are according to our measures.[6] The vertical axis is the fraction of the classrooms that have unusually large test score increases one year followed by especially small gains the next year. The graph combines all classrooms and all subjects in our data.[7] Consistent with our assumptions, for most of the range, there is virtually no relationship between how suspicious a classroom's answer strings are and the likelihood of large test score fluctuations. As one approaches the extreme right tail of the distribution of suspicious answer strings, however, the probability of large test score fluctuations rises dramatically, consistent with our conjecture that cheating classrooms should be extreme on both of our measures. To estimate the prevalence of

[6] We defer a precise discussion of how we construct our cheating indicators to Section III.

[7] To construct the figure, classes were rank ordered according to their answer strings and divided into 200 equally-sized segments. The circles in the figure represent these 200 local means. The line displayed in the graph is the fitted value of a regression with a seventh-order polynomial in a classroom's rank on the suspicious strings measure.

cheating, we will essentially compare the actual area under the curve in the far right tail of Figure 1 to the predicted area under the curve in the right tail under our maintained assumptions about how the two measures co-vary in non-cheating classrooms. Essentially, we predict that in the absence of cheating, the relationship between suspicious answer strings and large test score fluctuations would be linear over the entire range, rather than rising sharply in the tail. That sharp rise, we argue, is a consequence of cheating.

Empirically, we find evidence of cheating in approximately 200 classrooms per year in our data, or four to five percent of the classes in our sample. This estimate is likely to be a lower bound on the true incidence of cheating for two reasons. First, we focus only on the most egregious type of cheating, where teachers systematically altering student test forms. There are other more subtle ways in which teachers can cheat, such as providing extra time to students, that our algorithm is unlikely to detect. Second, even when test forms are altered, our approach is only partially successful in detecting illicit behavior. When we ourselves simulate cheating by altering student answer strings and then testing for cheating in the artificially manipulated classrooms, many instances of moderate cheating go undetected. This is particularly true if a teacher employs a limited amount of sophistication in the cheating (e.g. avoiding changing large blocks of consecutive questions for many students).

A number of patterns in the results reinforce our confidence that what we measure is indeed cheating. First, cheating on one part of the test (e.g., math) is a strong predictor of cheating on other sections of the test (e.g., reading). Second, cheating is correlated within classrooms over time and across classrooms in a particular school. Third, the students in classrooms with large test score gains that are most likely attributable to cheating lose most of

their gains the following year. In contrast, students in classrooms with large test-score gains that do not have suspicious answer string patterns retain the majority of their gains the next year, despite some loss due to mean reversion. Fourth, simulation results demonstrate that there is nothing mechanical about our identification approach that automatically generates patterns like those observed in the data. When we randomly assign students to classrooms and search for cheating in these simulated classes, our methods find little evidence of cheating. Finally, there is no evidence that we are mistaking teachers focusing effort on specific subject areas (e.g. algebra, fractions) for cheating. The classrooms we label as cheaters are no more likely to have their most suspicious answers cluster within a single topic than other classes.

In addition, the prevalence of cheating appears to respond to relatively minor changes in teacher incentives. The importance of standardized tests in the Chicago Public Schools increased substantially with a change in leadership in 1996. Schools that scored low on reading tests were placed on probation and faced the threat of reconstitution (although no elementary school has actually been reconstituted). In addition, students in certain grades were required to meet minimum test scores cutoffs in math and reading in order to advance to the next grade. Following the introduction of these policies, the prevalence of cheating rose sharply in classrooms with large numbers of low-achieving students. In contrast, classrooms with average or higher-achieving students showed no increase in cheating. Finally, cheating prevalence appears to be systematically lower in cases where the costs of cheating are higher (e.g. in mixed-grade classrooms in which two different exams are administered simultaneously) or the benefits of cheating are lower (e.g. in classrooms with more special education or bilingual students who take the standardized tests, but whose scores are excluded from official calculations).

The remainder of the paper is structured as follows. Section II presents a simple statistical model for detecting teacher cheating. Section III introduces the particular indicators we employ for detecting teacher cheating. Section IV provides a brief overview of the institutional details of the Chicago Public Schools and the data set that we use. Section V presents the basic empirical results on the prevalence of cheating. Section VI analyzes in greater detail the factors that influence which teachers cheat and for which students within the classroom the teachers cheat. Section VII discusses the results and the implications for increasing reliance on high-stakes testing. The appendix provides precise details of the construction of the cheating indicators used in the analysis.

II. A Statistical Model of Teacher Cheating

Assume that we have two measures of a classroom's outcome on a standardized test. $SCORE_c$ captures how well class c scores on the test, relative to how the same students have done on past standardized tests and will do on future tests. $ANSWERS_c$ measures how unusual are the pattern of answers given by students in class c (e.g. are there are unusual blocks of answers, or an especially high degree of correlation across student responses). For simplicity in presenting the model, we assume that these two measures take on one of two values: $SCORE_c$ = {low,high} and $ANSWERS_c$ = {typical,unusual}. Further suppose there are two types of classrooms: those in which teachers cheat, and those in which they do not. Define $CHEAT_c$ equal to one if cheating occurs, and zero otherwise. Our first critical assumption is as follows:

(A1) Had cheating classrooms not cheated, their distribution of the two outcome measures, SCORE and ANSWERS, would be identical to that of non-cheating classrooms.

Second, we assume that although cheating behavior is not directly observed, cheating increases the probability that a classroom will have a high average test score and an unusual pattern of answer strings:

$$\Pr(SCORE_c = high \mid CHEAT_c = 1) > \Pr(SCORE_c = high \mid CHEAT_c = 0) \quad (A2)$$

$$\Pr(ANSWERS_c = unusual \mid CHEAT_c = 1) > \Pr(ANSWERS_c = unusual \mid CHEAT_c = 0)$$

We define S_{nc} as the probability that a non-cheating class has a high value of *SCORE* and A_{nc} as the probability that a non-cheating class has an unusual value for *ANSWERS*. For purposes of exposition, let us assume that for non-cheating classrooms, the two measures *SCORE* and *ANSWERS* are uncorrelated (although this assumption will be relaxed in the empirical work). It then follows that:

$$S_{nc} \equiv \Pr(SCORE_c = high \mid CHEAT_c = 0, ANSWERS_c = typical) = \Pr(SCORE_c = high \mid CHEAT_c = 0, ANSWERS_c = unusual) \quad (A3)$$

$$A_{nc} \equiv \Pr(ANSWERS_c = unusual \mid CHEAT_c = 0, SCORE_c = low) = \Pr(ANSWERS_c = unusual \mid CHEAT_c = 0, SCORE_c = high)$$

The following Lemma follows directly from assumptions (A1) – (A3):

Lemma 1: Let $\bar{S}_{nc} = \Pr(SCORE_c = high \mid ANSWERS_c = typical)$ and let $\bar{A}_{nc} = \Pr(ANSWERS_c = unusual \mid SCORE_c = low)$, then $\bar{S}_{nc} \geq S_{nc}$ and $\bar{A}_{nc} \geq A_{nc}$.

Lemma 1 says that the average fraction of high test scores among classes with typical answer strings provides an upper bound on the probability that non-cheating classrooms will have high test scores. Similarly, the observed fraction of unusual answer strings among classes with low

test score fluctuations is an upper bound on the probability that non-cheating classrooms will have unusual answer strings. The reason these values are upper bounds is because some classrooms that have "low" test scores or "typical" answer strings may actually be cheaters that our methods fail to detect. Even if cheaters are low on one dimension (either *SCORE* or *ANSWERS*), they are still more likely to be elevated on the other measure, leading us to overstate the baseline rate of high test scores or unusual answer strings among non-cheaters. Only if all cheating classrooms have high test scores and unusual answer strings will the bounds in Lemma 1 be strict.

Denote the total number of classrooms as N and the total number of classrooms that have both high test scores and unusual answer strings as N_{hu}. Then, a lower bound on the number of cheating classrooms is how many extra rooms there are with both high test scores and unusual answer strings, relative to the number that would be expected if no classrooms cheated:

$$\hat{N}_{cheat} = (N_{hu} - N \times \bar{S}_{nc} \times \bar{A}_{nc}) \quad (1)$$

$\hat{N}_{cheat}$ represents a lower bound on the number of cheating classrooms for two reasons. First, some cheating classrooms will not be detected by our measures and so will not register as having high test scores and unusual strings. Second, by Lemma 1, the probabilities of high test scores or unusual answer strings among non-cheating classes are upper bounds on the true values.

Calculations like those in equation (1) provide the basis for our estimation of the number of cheating classrooms. In our empirical work, we generalize (1) by allowing for correlation between test scores and answer string patterns in non-cheating classrooms, but the logic is unchanged.

One important caveat to note is that we cannot identify any individual classroom as cheating or not cheating with perfect certainty. The probability that a class with high test score fluctuations and unusual answer strings is cheating is given by:

$$(2) \qquad \Pr(CHEAT_c = 1 \mid SCORE_c = high, ANSWERS_c = unusual) = \frac{N_{cheat}}{N_{hu}}$$

As the thresholds for what constitutes a "high" test score or an "unusual" answer strings are made more stringent, N_{hu} will decline and, consequently, our level of certainty rises that any particular classroom exhibiting these characteristics is cheating. In essence, raising these thresholds will decrease the number of false positives in our estimates.

III. Indicators of Teacher Cheating

We employ two types of measures to detect cheating. One indicator captures predictable fluctuations in test scores that are likely to be associated with cheating. The other indicator summarizes the extent to which answer strings in a classroom appear unusual or suspicious. In this section, we discuss informally the indicators we use to detect cheating, and then provide a concrete example that compares data from two actual classrooms: one in which there appears to be cheating and one in which there does not. Readers interested in a more rigorous description of how the indicators are constructed are directed to the Appendix.

In selecting our measures of cheating, we focus on detecting teacher actions that lead to large, artificial increases in test scores for a large number of students in the class.[8] By focusing

[8] We have no way of knowing whether the patterns we observe arise because a teacher explicitly alters students' answer sheets, directly provides answers to students during a test, or perhaps makes test materials available to students in advance of the exam (for instance, by teaching a reading passage that is on the test). If we had access to the actual exams, it might be possible to distinguish between these scenarios through an analysis of erasure patterns.

on the entire classroom, we are unlikely to misclassify cheating by individual students as teacher cheating. Teacher actions such as "teaching to the test" or allowing students extra time to complete exams are not likely to be detected by our measures because they are unlikely to generate sufficiently unusual response patterns in the answer strings.

Within this already restrictive definition of teacher cheating, we narrow our focus even further by excluding a particular form of cheating that appears to be quite prevalent in the data: teachers randomly filling in answers left blank by students. For example, in some classrooms, almost every student will end the test with a long string of "B's" or an alternating pattern of "B" and "C." The fact that almost all students in the class coordinate on the same pattern strongly suggests that the students themselves did not fill in the blanks, or were under explicit instructions by the teacher to do so. Since there is no penalty for guessing on the test, filling in the blanks can only increase student test scores. While this type of teacher behavior is likely to be viewed by many as unethical, we do not make it the focus of our analysis because (1) it is difficult to provide definitive evidence of such behavior (a teacher could argue that he or she instructed students well in advance of the test to fill in all blanks with the letter "C" as part of good test-taking strategy), and (2) in our minds it is categorically different than a teacher who systematically changes student responses to the correct answer.

Cheating Indicator #1: Unexpected Test Score Fluctuations

Given that the aim of cheating is to raise test scores, an obvious potential indicator of teacher cheating is a classroom that experiences unexpectedly large gains in test scores relative to how those same students tested in the previous year. Since test score gains that result from

cheating do not represent real gains in knowledge, there is no reason to expect the gains to be sustained on future exams taken by these students (unless, of course, next year's teachers also cheat on behalf of the students). Thus, large gains due to cheating should be followed by smaller than usual test score gains for these students in the following year. In contrast, if large test score gains are due to a talented teacher, the student gains are likely to have a greater permanent component, even if some regression to the mean occurs.

In practice, the choice of a cutoff for what represents an "unexpectedly" large test score gain or loss is somewhat arbitrary. Our admittedly simple approach is to rank each classroom's average test score gains relative to all other classrooms in that same subject, grade, and year,[9] and construct the following statistic:

$$(3) \qquad SCORE_{cbt} = (rank\_gain_{c,b,t})^2 + (1 - rank\_gain_{c,b,t+1})^2$$

where $rank\_gain_{cbt}$ is the percentile rank for class c in subject b in year t. Classes with relatively big gains on this year's test and relatively small gains on next year's test will have high values of *SCORE*. Squaring the individual terms gives more relatively more weight to big test score gains this year and big test score declines the following year.[10] In the empirical analysis, we consider three possible cutoffs for what it means to have a "high" value on SCORE, corresponding to the 80th, 90th, and 95th percentiles among all classrooms in the sample.

[9] We also experimented with more complicated mechanisms for defining large or small test score gains (e.g., predicting each student's expected test score gain as a function of past test scores and background characteristics and computing a deviation measure for each student which was then aggregated to the classroom level), but because the results were similar we elected to use the simpler method. We have also defined gains and losses using an absolute metric (e.g., where gains in excess of 1.5 or 2 grade equivalents are considered unusually large), and obtain comparable results.

[10] In the following year the students who were in a particular classroom are typically scattered across multiple classrooms. We base all calculations off of the composition of this year's class.

Cheating Indicator #2: Suspicious Answer Strings

Teacher cheating, particularly if accomplished by the teacher actually changing answers on test forms, is likely to leave a discernible trail in student answer strings. The quickest and easiest way for a teacher to cheat is to alter the same block of consecutive questions for a substantial portion of students in the class. More sophisticated interventions might involve skipping some questions so as to avoid a large block of identical answers, or altering different blocks of questions for different students.

We combine four different measures of how suspicious a classroom's answer strings are in determining whether a classroom may be cheating. The first measure focuses on the most unlikely block of *identical* answers given by students on consecutive questions. Using past test scores, future test scores, and background characteristics, we predict the likelihood that each student will give each possible answer (A, B, C or D) on every question using a multinomial logit. This means that each student's predicted probability of choosing a particular response is identified by the likelihood that other students (in the same year, grade and subject) with similar background characteristics will choose that response. We then search over all combinations of students and consecutive questions to find the block of identical answers given by students in a classroom least likely to have arisen by chance.[11] The more unusual is the most unusual block of test responses (adjusting for class size and the number of questions on the exam, both of which increase the possible combinations over which we search), the more likely it is that cheating

[11] Note that we do not require the answers to be correct. Indeed, in many classrooms, the most unusual strings include some incorrect answers. Note also that these calculations are done under the assumption that a given student's answers are uncorrelated (conditional on observables) across questions on the exam, and that answers are uncorrelated across students. Of course, this assumption is unlikely to be true. Since all of our comparisons rely on the *relative* unusualness of the answers given in different classrooms, this simplifying assumption is not problematic unless the correlation within and across students varies by classroom.

occurred. Thus, if ten very bright students in a class of thirty give the correct answers to the first five questions on the exam (typically the easier questions), the block of identical answers will not appear unusual. In contrast, if all fifteen students in a low-achieving classroom give the same correct answers to the last five questions on the exam (typically the harder questions), this would appear quite suspect.

The second measure of suspicious answer strings involves the overall degree of correlation in student answers across the test. When a teacher changes answers on test forms, it presumably increases the uniformity of student test forms across students in the class. This measure is meant to capture more general patterns of similarity in student responses that goes beyond just identical blocks of answers. Based on the results of the multinomial logit described above, for each question and each student we create a measure of how unexpected the student's response was. We then combine the information for each student in the classroom to create something akin to the within-classroom correlation in student responses. This measure will be high if students in a classroom tend to give the same answers on many questions, especially if the answers given are unexpected, i.e. correct answers on hard questions or systematic mistakes on easy questions.

Of course, within-classroom correlation may arise for many reasons other than cheating (e.g., the teacher may emphasize certain topics during the school year). Therefore, a third indicator of potential cheating is a high *variance* in the degree of correlation *across* questions. That is, on some questions students' answers are highly correlated, but on other questions they are not. If the teacher changes answers for multiple students on selected questions, the within-class correlation on those particular questions will be extremely high, while the degree of within-

class correlation on other questions is likely to be typical. This leads the cross-question variance in correlations to be larger than normal in cheating classrooms.

Our final indicator compares the answers that students in one classroom give compared to other students in the system who take the identical test and get the exact same score. Questions vary significantly in difficulty. The typical student will answer most of the easy questions correctly and get most of the hard questions wrong (where "easy" and "hard" are based on how well students of similar ability do on the question). If students in a class systematically miss the easy questions while correctly answering the hard questions, this may be an indication of cheating.

Our overall measure of suspicious answer strings is constructed in a manner parallel to our measure of unusual test score fluctuations. Within a given subject, grade, and year, we rank classrooms on each of these four indicators, and then take the sum of squared ranks across the four measures:[12]

$$(4) \qquad ANSWERS_{cbt} = (rank\_m1_{c,b,t})^2 + (rank\_m2_{c,b,t})^2 + (rank\_m3_{c,b,t})^2 + (rank\_m4_{c,b,t})^2$$

In the empirical work, we again use three possible cutoffs for potential cheating: 80th, 90th, and 95th percentiles.

A comparison of two actual classrooms

Figure 2, which presents student answer strings test scores for two classrooms, provides an example of how our cheating indicators work in practice. The top panel of data is a class in which we suspect teacher cheating took place; the bottom panel corresponds to a typical

classroom. Each row in Figure 2 represents one student's answers to each item on the test. Columns correspond to the different questions asked. The letter "A," "B," "C," or "D" means a student provided the correct answer. If a number is entered, the student answered the question incorrectly, with "1" corresponding to a wrong answer of "A," "2" corresponding to a wrong answer of "B," etc. On the right-hand side of the table, we also present student test scores for the preceding, current, and following year.

Focusing first on the patterns in the answer strings in Figure 2, we see that for most of the test, correct and incorrect answers are sporadically interspersed with no discernible pattern. In the top panel of the figure, however, over half of the students in the class provide the same answers to nine consecutive questions towards the end of the test, suggesting teacher cheating. That places the classroom in the 99th percentile on our first measure of suspicious strings which focuses on blocks of identical answers. In the bottom panel, the most unusual block of answers is the "2CAD2C" given by two students (the fifth and sixth students listed) starting on question 23 of the test. The bottom classroom ranks in the 43rd percentile on this measure.

Questions on which students in our two sample classrooms have somewhat elevated within-class correlations are demarcated by an "*" at the bottom of the column corresponding to that question. For cases of extreme correlations, an "!" is given. Not surprisingly, the correlation is very high on the questions that are part of the suspicious string in the class suspected of cheating. It is also somewhat elevated on the block of questions with similar answers in the non-cheating classroom. For the remaining parts of the exam, the indicator is similar across the two classrooms. The within-class degree of correlation in the top classroom

[12] Because different subjects and grades have differing numbers of questions, it is difficult to make meaningful

places it at the 99th percentile among classrooms. The bottom class, in contrast, is in the 49th percentile on this measure. With respect to the variance in within-class correlation across questions, the top class is at the 99th percentile; the bottom class is at the 32nd.

Although not shown directly in the table, the top classroom also fares poorly on our last measure of suspicious strings – the degree to which students in this class tend to get the same questions right and wrong as students in other classes. Because the questions near the end of the test are difficult (note how few of the students in the second class get these questions correct), students in this first class look very unusual relative to other students in the system. The class once again ranks near the 99th percentile on this measure, compared to only the 17th percentile for the class in the bottom panel.

Turning our attention to the test scores on the right-hand side of Figure 2, mean test scores in the previous year are similar for the two classes. On that year's test, however, the top classroom suspected of cheating experienced an enormous jump in test scores (1.7 grade equivalents on average, compared to a mean of 0.9 for all classrooms in this subject, grade, and year). The bottom classroom had a typical gain. In the following year, students in the top class actually see test score *declines* on average, whereas students in the bottom panel continue to progress at a normal rate. Note also that it is only the students in the top panel who are part of the unusual answer strings that exhibit enormous test score gains followed by large declines. Among the handful of students in the top panel that do not appear to have been the beneficiaries of the cheating, the test score gains in the current and following year are typical. The classroom

---

comparisons across tests on the raw indicators.

in the bottom part of Figure 2 would not qualify as having unusual test score fluctuations by any of our cutoffs; the top classroom qualifies on even the strictest definition.

IV. Data and Institutional Background

Elementary students in Chicago public schools take a standardized, multiple-choice achievement exam known as the Iowa Test of Basic Skills (ITBS). The ITBS is a national, norm-referenced exam with a reading comprehension section and three separate math sections.[13] Third through eighth grade students in Chicago are required to take the exams each year. Most schools administer the exams to first and second grade students as well, although this is not a district mandate.

Our base sample includes all students in third to seventh grade for the years 1993.[14] For each student, we have the question-by-question answer string on each year's ITBS reading comprehension and mathematics tests, school and classroom identifiers, the full history of prior and future test scores, and demographic variables including age, sex, race, and free lunch eligibility. We also have information about school-level characteristics including mobility, poverty and attendance rates, racial composition and average teacher characteristics including percent with an MA+ degree, years of experience, and undergraduate major. We do not, however, have individual teacher identifiers, so we are unable to directly link teachers to classrooms or to track a particular teacher over time.

[13] There are also other parts of the test which are either not included in official school reporting (spelling, punctuation, grammar) or are given only in select grades (science and social studies), for which we do not have information.

[14] We exclude eighth graders because our algorithm requires test score data for the following year and the ITBS test is not administered to ninth graders. Another standardized test is given to ninth graders, but a substantial fraction of the students fail to take that test and it is not directly comparable to the elementary exams.

Because our cheating proxies rely on comparisons to past and future test scores, we drop observations that are missing reading or math scores in either the preceding year or the following year.[15] Students with missing demographic data are also excluded from the analysis. Finally, because our algorithms for identifying cheating rely on identifying suspicious patterns within a classroom, our methods have little power in classrooms with small numbers of students. Consequently, we drop all classrooms for which we have fewer than ten valid students in a particular grade after our other exclusions. A handful of classrooms with impossibly large number of students – presumably multiple classrooms combined into one – are also dropped. Our final data set contains roughly 20,000 students per grade per year distributed across approximately 1,000 classrooms, for a total of over 40,000 classroom-years of data (with four subject tests per classroom-year) and over 700,000 student-year observations.

Summary statistics for the full sample of classrooms are shown in the first column of Table 1, with the unit of observation being at the level of class*subject*year. The second column of Table 1 reports student-level summary statistics for the subset of our sample of classrooms we classify as likely to have cheated. Classrooms we suspect of cheating are more likely to be subject to the accountability policies and students in these classes are disproportionately from the bottom half of the achievement. School-level teacher characteristics do not differ much between the whole sample and the suspected cheaters.

The ITBS exams are administered over a week long period in early May. Third grade teachers are permitted to administer the exam to their own students, while other teachers switch classes to administer the exams. The exams are generally delivered to the schools one to two

[15] Test data may be missing either because a student did not attend school on the days of the test, or because the

weeks before testing, and are supposed to be kept in a secure location by the principal or the school's test coordinator, an individual in the school designated to coordinate the testing process (often a counselor or administrator). Each section of the exam consist of 30 to 60 multiple choice questions which students are given between 30 and 75 minutes to complete.[16] Students mark their responses on answer sheets, which are scanned to determine a student's score. Teachers or administrators then "clean" the answer keys, erasing stray pencil marks, removing dirt or debris from the form, and darkening item responses that were only faintly marked by the student. At the end of the week, the test coordinators at each school deliver the completed answer keys and exams to the CPS central office. School personnel are not permitted to keep copies of the actual exams, although school officials acknowledge that a number of teachers each year do so. The CPS has administered three different versions of the ITBS between 1993 and 2000. The CPS alternates forms each year, with new forms being offered for the first time in 1993, 1994 and 1997.[17]

The exams are scored electronically by CPS central office staff. There is no penalty for guessing, so that a student's raw score is simply calculated as the sum of correct responses on the exam. The raw score is then translated into a metric known as grade equivalents, which are normed so that a student at the 50th percentile in the nation scores at the eighth month of her

---

student transferred into the CPS system in the current year or left the system prior to the next year of testing.

[16] The mathematics and reading tests measure basic skills. The reading comprehension exam consists of three to eight short passages followed by up to nine questions relating to the passage. The passages include poetry, fictional stories, and narratives on historical, scientific or literary topics. The questions assess factual recall (e.g., Who was the main character in the story?) as well as critical analysis (e.g., What was the main idea of the passage?) and interpretation (e.g., How do you think Jose felt at the end of the story?). The math exam consists of three sections that assess number concepts, problem-solving and computation.

[17] These three forms are used for re-testing, summer school testing, and mid-year testing as well, so that it is likely that over the years, teachers have seen the same exam on a number of occasions.

current grade. For example, an average third grader taking the test in the eighth month of third grade will score a 3.8. Similarly, a sixth grader that scores a 5.8 is one year behind grade level.

V. Estimating the Prevalence of Cheating

We undertake a determination of the prevalence of cheating in three steps. First, we analyze whether our cheating measures are actually effective at detecting cheating. To do this, we artificially alter student answers in ways that might mimic teacher cheating and then see if we catch the classrooms for which we ourselves have cheated. In addition, we also alter student answers in a way that may reflect a class that has large, legitimate test score gains due to an outstanding teacher, demonstrating that our measures are much less likely to mistakenly classify such gains as cheating. Second, we examine the empirical distribution of our cheating measures in the portion of the sample unlikely to be heavily contaminated by cheating. This provides the basis for constructing a non-cheating counterfactual against which the actual data is compared. Finally, we present the basic findings with respect to the estimated cheating rate.

a) Do the cheating measures actually detect cheating classrooms?

Because we do not know which classrooms are cheating, we have no direct way of knowing whether our measures actually detect cheating. One indirect way of testing this hypothesis, however, is to simulate cheating and then ascertain whether our measures detect this artificial cheating. In this section, we simulate two different types of teacher cheating. The first is a very naive version, in which a teacher starts cheating at the same question for a number of students and changes consecutive questions to the right answers for these students, creating a

block of identical and correct responses. The second type of cheating is much more sophisticated: we *randomly* change answers from incorrect to correct for selected students in a class.[18] For the purposes of comparison, we also present results attempting to simulate the effects of a good teacher inducing the same gain among students in the class. The impact of a good teacher will differ in two ways from a cheating teacher: (1) some of the gains will be preserved in the following year, and (2) the students will not get random answers correct, but rather, will tend to show the greatest improvement on the easiest questions that the students were getting wrong. For both types of cheating and the good teacher scenario, we run simulations changing 3, 6 or 9 questions for 25, 50 or 100 percent of the students in the classroom. We alter the answer strings for every classroom, one classroom at a time, tallying the fraction of the cases in which the artificially cheating classroom exceeds our strictest threshold (95th percentile) on both *ANSWERS* and *SCORE*. We present the results for 5th grade reading in 1993. This grade and year was selected because we want our baseline sample to be as free of cheating as possible. In theory, the incentives for teachers to cheat in that grade and year are low because this is before accountability reforms and fifth-grade test outcomes were not widely publicized at that time.

Table 2 reports the results of the simulation. As a point of reference, 1.13 percent of the classrooms in the actual data exceed the thresholds we use in the table for labeling a classroom as cheating. For the most minor case of cheating (3 questions for 25 percent of the class), our cheating indicator picks up less than 4 and 2 percent respectively of the unsophisticated and sophisticated cheating. The good teacher is no more likely to be labeled a cheater in this case than is a randomly drawn classroom in the actual data. As the extent of cheating increases –

[18] We have also experimented with forms of cheating with intermediate degrees of sophistication. Not surprisingly,

either by increasing the number of students or the number of questions altered – the success of the algorithm improves greatly, but it is still far from perfect. If six questions are altered for half the class, more than 50 percent of the unsophisticated cheaters are detected, as well as more than one-third of the sophisticated cheaters. The likelihood that a good teacher is labeled as a cheater is still less than 3 percent. By the most extreme cases we examine (9 questions for 100 percent of the class), roughly 90 percent of classrooms are categorized as cheating. In the case, good teachers will be identified as cheating nearly 40 percent of the time as well. Thus, to the extent that there are teachers capable of such remarkable feats (it implies that the mean test score gain in the classroom in one year is well over two grade-equivalents, something observed in roughly one in 1,000 classrooms in our sample, even with cheaters included), there is a substantial likelihood we will mistakenly label them as cheaters.[19]

In summary, our cheating indicators are quite effective at detecting extreme instances of cheating, even if done in a sophisticated manner by the teacher. Many more limited cases of cheating will not be detected by our measures, particularly if the cheating is done cleverly. Thus, to the extent that actual cheating done by teachers is moderate in degree and/or of a sophisticated kind, many cheaters will slip through the cracks, and our estimates of the prevalence of cheating classrooms are likely to be (perhaps very loose) lower bounds on the true values. On the other hand, our algorithm does yield some false positives, which works in the opposite direction.

---

the effectiveness of our measures in detecting moderately sophisticated types of cheating falls in between our ability to detect cheating in the two polar cases we present.

[19] To the extent this is a major concern (e.g., if these results were going to be used in disciplinary actions), there are alternative measures that could be employed which are less likely to catch actual cheaters, but also dramatically reduce the likelihood that a good teacher would falsely be accused of cheating. One such measure would be to require that most or all of the current year's gain is lost in the following year.

b) Projecting the distribution of the cheating measures in a counterfactual with no cheating

If we knew with certainty the classrooms that were cheating, then the distribution of our cheating measures in non-cheating classrooms would be directly observable.[20] Our sample, however, is made up of an unknown mixture of cheating and non-cheating classes. Since cheating classrooms are likely to have large test score fluctuations and unusual answer strings, they are likely to be concentrated in the upper tails, leaving the remainder of the distribution relatively free of cheating. By observing patterns in that portion of the distribution, it is possible to make reasonable predictions as to what the upper tail might look like absent cheating.[21]

The top panel of Table 3 presents breakdowns of the fraction of cases in which *SCORE* is above each of our cutoffs, as a function of the quartile that *ANSWERS* falls into, omitting the top quartile of *ANSWERS* because that is where the cheating classrooms are likely to be. Among classrooms with *ANSWERS* are in the 0-25th percentile (i.e. not suspicious at all), 16.3 percent have test score gains above the 80th percentile, 6.2 percent are above the 90th percentile, and 2.3 percent are above the 95th percentile. As one moves to the right in the table, the answer strings are becoming more suspicious. The frequency of high values of *SCORE* rises slightly moving to the right, but the relationship is weak. This mirrors the pattern presented earlier in Figure 1.

---

[20] Of course, if we knew which classrooms were cheating, then this assumption would not be necessary in the first place.

[21] There are two potential biases at work in this sort of analysis. First, some cheating classrooms are likely to evade our measures and slip into the sample we are describing as non-cheaters. Cheating classrooms will probably be especially prevalent in the 50-75th quartile – classes that are somewhat above average in terms of suspiciousness. Thus, one might observe a spurious rise in the frequency of extreme values as one moves from classrooms that are not all suspicious to those that are somewhat suspicious, simply because the fraction of undetected cheaters rises. On the other hand, it may be the case that the degree of correlation between *ANSWERS* and *SCORE* may vary over the distribution. In particular, there may be a positive correlation between those two variables in the right-tail of the distribution, even if teachers are not cheating. This would lead us to overstate the number of cheaters.

The bottom panel of Table 3 reverses the exercise, showing how the likelihood of having suspicious answer strings changes as one moves from low quartiles of *SCORE* to high quartiles. Note that the patterns are reversed. Classes that are in the bottom quartile on *SCORE* have relatively *high* rates of suspicious strings. The probabilities of suspicious strings for classes in the second and third quartile are nearly identical to one another, and much lower than those of the first quartile.

The key question here is the frequency with which one would observe extreme values of one of the cheating measures when the other cheating measure is also an outlier, assuming no cheating occurred. One can imagine a variety of sensible approaches to predicting the upper tail using the information in Table 3. For example, one could fit a linear or quadratic model to the patterns observed in the lower portion of the distribution and extrapolate the estimates. In practice, however, the trends in the 0-75th percentile of the distribution are so weak that the results we obtain are not sensitive to the precise formulation. Consequently, we simply use the values in the 50-75th percentile as our estimate of what would have happened in the upper tail, absent cheating.

c) Estimating the Prevalence of Cheating

Our key equation for estimating the frequency of cheating, presented earlier in the modeling section, is

(1) $$\hat{N}_{cheat} = (N_{hit} - N \times \bar{S}_{nc} \times \bar{A}_{nc})$$

which simply says that the estimated number of cheating classes is equal to the number of classes that are outliers on both the cheating measures, minus the number of classes we would expect by chance to be high on both measures based on the distribution of the measures in non-cheating classrooms. Guided by the results presented above, we compute $\bar{S}_{nc}$ as the fraction of classes with *SCORE* above a certain threshold, conditional on being in the 50-75th percentile on answers, and vice-versa for $\bar{A}_{nc}$.

The top panel of Table 4 presents our estimates of the percentage of classrooms that are cheating on average on a given subject test (i.e. reading comprehension or one of the three math tests) in a given year. We present a 3 x 3 matrix of estimates corresponding to how stringent the thresholds are for judging whether a classroom's test score fluctuations and answer string patterns qualify as suspicious. The estimated prevalence of cheaters ranges from 1.1 percent to 2.1 percent, depending on the particular set of cutoffs used. As would be expected, the number of cheaters is generally declining as higher thresholds are employed. Nonetheless, it is encouraging that over such a wide range of cutoffs, the range of estimates is relatively tight.

The bottom panel of Table 4 presents estimates of the percentage of classrooms that are cheating on *any* of the four subject tests in a particular year. If every classroom that cheated did so only on one subject test, than the results in the bottom panel would simply be four times the results in the top panel. In many instances, however, classrooms appear to cheat on multiple subjects. Thus, the prevalence rates range from 3.4-5.6 percent of all classrooms.[22]

[22] Computation of the overall prevalence is relatively complicated because it involves calculating not only how many classrooms are actually above the thresholds on multiple subject tests, but also how frequently this would occur in the absence of cheating. The full programming solution to this problem is available from the authors.

Table 5 presents breakdowns of cheating by grade and subject. Rows in the table correspond to grades and columns represent different subject tests. Several interesting findings stand out. The third grade has the highest cheating rates, which is likely due to the fact that this grade is the only one in which teachers are permitted to administer the exams to their own students. Cheating rates also tend to be higher in sixth grade, which may be due to the fact that this grade has traditionally been a focus of greater attention on the part of teachers and administrators because statewide achievement exams are also administered in this grade. If one looks at specific subject tests, cheating rates are slightly higher on reading comprehension and the first and third math exams, while noticeably lower on the second math exam. This may be because the second math exam is located in the middle of the answer key, making it more difficult to quickly change student answers. In addition, in the upper grades, this section consists largely of questions that ask students to interpret graphs, charts and tables, and most students do relatively well on this section, so it is possible that teachers do not feel that they need to artificially inflate scores in this area.

VI. Are We Really Detecting Cheating?

If accurate, the results above suggest cheating rates of 4-5 percent among Chicago elementary school classrooms. Because of the necessarily indirect nature of our identification strategy, a healthy skepticism towards our conclusions may be warranted. In this section, we provide a range of supplemental analyses suggesting that the results are indeed cheating and addressing possible competing explanations as to why the patterns we observe may have arisen.

a) Will our method find cheating, even if no cheating exists?

Our identification relies on assumptions about how variables will be distributed in the tails of a distribution that is not directly observed. It is conceivable that our assumptions are inaccurate, falsely generating what appears to be evidence of cheating, even when no such cheating occurred.

As a test of this possibility, we randomly assigned students to hypothetical classrooms. These synthetic classrooms thus consisted of groups of students who in actuality had no connection to one another. We then analyzed these hypothetical classrooms using the same algorithm applied to the actual data. As one would hope, no evidence of cheating was found in the simulated classes. Indeed, the estimated prevalence of cheating was slightly negatve in this simulation, i.e. classrooms with large test score increases in the current year followed by big declines the next year were slightly *less* likely to have unusual patterns of answer strings. Thus, we conclude that there is no evidence that our identification approach finds cheating even when no cheating is actually present.

b) Are classrooms with suspicious answer strings less likely to maintain large test score gains?

Test score gains due to cheating should be completely transitory, assuming that the likelihood of having a cheating teacher next year is the same for students who do or do not have a cheating teacher this year. In contrast, while there might be substantial mean reversion for classrooms with large test score gains for reasons other than cheating, there might also be a permanent component to such gains.

In conducting such a hypothesis test, we cannot, of course, use information about next year's test score as a basis for labeling a classroom a likely cheater (as we generally do in our cheating classification, see equation (2)). We can, however, compare classrooms with large test score gains this year that either do or do not have suspicious patterns of answer strings. The more suspicious the answer strings, the more likely is cheating, and the greater the expected mean reversion in the following year's test scores.

Table 6 presents the results of this analysis. The top panel of the table restricts the sample to the ten percent of classrooms with the greatest test score gains in reading in the current year, relative to other classes in that grade and year. The middle panel and bottom panel look at the top 5 percent and top 1 percent respectively of classroom test score gains.[23] Reported in the table are the mean excess gain in these classes in the base year (i.e., the gains above and beyond the mean gain for all classrooms in the system on a given subject, grade, and test that year), the excess gain for the same students the following year, and the percent of the base year excess gain that is maintained. Columns in the table correspond to how suspicious the classroom's answer string patterns were: below the 50th percentile, 50th-80th percentile, 80th-95th percentile, 95th-99th, and greater than the 99th percentile. We expect the fraction of cheating classrooms to increase as the answer strings become more suspicious. Consequently, the fraction of the current year's excess test score gain relative to the mean classroom in that subject, grade, and year that is maintained the following year should decrease moving from left to right in the table.

Looking first at the top panel of Table 6, among the ten percent of classes with the greatest test score gains, those whose answer strings are not at all suspicious (below the 50th

[23] Reiterating what was written above, it cannot be emphasized enough that this table differs from our previous

percentile on *ANSWERS*) gain on average .59 grade equivalents more than the system-wide mean in the base year. Students in these classrooms do perform slightly worse than expected the following year (-.11 grade equivalents). Nonetheless, at the end of the next year, these students maintain 81 percent of the base year excess gain through the following year (i.e., (.59-.11)/.59=.81). As one moves from left to right in the table towards classrooms more likely to be cheating, an increasing fraction of the current year's gains are lost, as predicted. In the most extreme cases of the one percent of the answer strings that are most suspicious, less than 13 percent of the apparent gains in the current year evaporate in the next year's test.[24] Note also that classes with suspicious answer strings are greatly over-represented among those achieving large test score gains, consistent with the prediction that our two cheating indicators will be highly positively correlated in the upper tail.

One possible concern that arises from the top panel of the table is whether the mean reversion is greater on the right-hand-side columns of the table simply because the base-year gains are larger. The bottom two panels of Table 6 demonstrate that this is not the case. When the sample is restricted to classrooms whose test score gains are in the top five percent or top one percent of all classes, very similar patterns appear. Classrooms that do not have suspicious answer strings continue to exhibit little mean reversion, even though the base year gains in this subset are even greater than the base year gains in the far right column of the top panel. The more suspicious the answer strings, the greater the mean reversion. Thus, these results are

analysis in that we are in no way conditioning on the following year's test scores, unlike when we construct our cheating estimates.

[24] We do not expect the test score gains to completely disappear because even among the classes with very suspicious answer strings, not all of the classrooms are cheating.

consistent with the hypothesis that cheating is the explanation for the large number of classes with both suspicious answer strings and large test score gains.

c) Do the same teachers and schools tend to cheat repeatedly?

If what we are detecting is truly cheating, then one would expect that a teacher who cheats on one part of the test would be more likely to cheat on other parts of the test. Also, a teacher who cheats one year would be more likely to cheat the following year. Finally, to the extent that cheating is either condoned by the principal or carried out by the test coordinator, one would expect to find multiple classes in a school cheating in any given year, and perhaps even that cheating in a school one year predicts cheating in future years. If what we are detecting is not cheating, then one would not necessarily expect to find strong correlation in our cheating indicators across exams for a specific classroom, across classrooms, or across years.[25]

Table 7 reports regression results testing these predictions. The dependent variable is an indicator for whether we believe a classroom is likely to be cheating on a particular subject test using our most stringent definition (above the 95th percentile on both cheating indicators). The baseline probability of qualifying as a cheater for this cutoff is 1.1 percent. To fully appreciate the enormity of the effects implied by the table, it is important to keep this very low baseline in mind. We report estimates from linear probability models (probits yield similar marginal effects), with standard errors clustered at the school level.

[25] Alternatively, if one thought that cheating were an individual teacher phenomenon, but school improvement, instructional quality or curricular content were a school-wide phenomena, then one might construe correlations within schools and over time as evidence against cheating. Given the fact that most teachers do not monitor their own exams, and that the test coordinator plays such a large role in the testing process within each school, we tend to think this scenario is less plausible.

Column 1 of Table 7 shows that cheating on other tests in the same year is an extremely powerful predictor of cheating in a different subject. If a classroom cheats on exactly one other subject test, the predicted probability of cheating on this test increases by over ten percentage points. Since the baseline cheating rates are only 1.1 percent, classrooms cheating on exactly one other test are ten times more likely to have cheated on this subject than are classrooms that did not cheat on any of the other subjects (which is the omitted category). Classrooms that cheat on two other subjects are almost 30 times more likely to cheat on this test, relative to those not cheating on other tests. If a class cheats on all three other subjects, it is 50 times more likely to also cheat on this test.

There also is evidence of correlation in cheating within schools. A ten percentage-point increase in cheating classrooms in a school (excluding the classroom in question) on the same subject test raises the likelihood this class cheats by roughly .016 percentage points. This potentially suggests some role for centralized cheating by a school counselor, test coordinator or the principal, rather than by teachers operating independently. There is little evidence that cheating rates within the school on other subject tests affects cheating on this test.

When making comparisons across years (columns 3 and 4), it is important to note that we do not actually have teacher identifiers. We do, however, know what classroom a student is assigned to. Thus, we can only compare the correlation between past and current cheating in a given *classroom*. To the extent that teacher turnover occurs or teachers switch classrooms, this proxy will be contaminated by serious measurement error. Even given this important limitation, cheating in the classroom last year predicts cheating this year. In column 3, for example, we see that classroom's that cheated in the same subject last year are 9.6 percentage points more likely

to cheat this year, even after we control for cheating on other subjects in the same year and cheating in other classes in the school. Column 4 shows that prior cheating in the school strongly predicts the likelihood that a classroom will cheat this year.

d) Is it really student cheating rather than teacher cheating that we are detecting?

Although our explanation focuses on teachers, it is also conceivable that there is something unusual about students in classes that we label as cheating (e.g., a particular set of student skills, high effort at the beginning of the test but not at the end, students copying off other student exams) that might lead to suspicious answer strings. If that is the case, then one would predict that the same students who have suspicious answer strings one year would also be likely to have suspicious answer strings the next year – much as we observed above that cheating in a classroom one year predicts cheating in a classroom the next year. We test this hypothesis in two ways. First, for each student in a given subject and year, we compute how unlikely it was for the most unusual block of answers that student was part of to occur (this corresponds to the first of the suspicious string measures we introduced in Section III). We then calculate the student-level correlation from one year to the next on that measure. The correlation is approximately .01, suggesting that students who are part of suspicious blocks of answers one year are not especially likely to be part of suspicious blocks the following year. Similarly, we compare the same student across years to determine whether some students tend to systematically get hard questions correct and easy questions wrong (the last of the suspicious string measures we introduced). The year-to-year correlation in that measure is .07. The absence in persistence in suspicious answer patterns over time for a given student makes it

unlikely that our results are being driven by student cheating, since one would expect that students who cheat one year would be likely to cheat with high probability the next year as well.

e) Are we mistaking emphasis on certain subject material for cheating?

If a math teacher spends several months on fractions with a particular class, one would expect the class to do particularly well on all of the math questions relating to fractions and perhaps worse than average on other math questions.[26] Such patterns might wrongly lead us to deem a classroom's answer strings to be suspicious.[27]

Table 8 shows OLS estimates of the relationship between whether a classroom is categorized as a cheater and the nature of the across-question correlations in the classroom on the first math exam. Seven different math skills/areas are tested in this exam: numeration, geometry, measurement, fractions, algebra, statistics and estimation. The dependent variable in the first and third columns is the number of different item-types for the three and five questions in a particular classroom that are most suspicious. The dependent variable in the second and fourth columns is a binary indicator of whether the three (five) most suspicious questions all fell in one area. Note first that the mean values, reported in brackets, show that there is generally very little concentration in the types of questions that are most suspicious within a classroom. For instance,

[26] One might imagine a similar scenario on the reading exam. If a teacher spends an entire semester studying the Underground Railroad, and the reading exam that year happens to include a passage on this topic, it would not be surprising to find that an extremely high number of students in the class correctly answer all of the items relating to this passage, which may appear as a highly suspicious answer string. However, it is also likely that when teachers cheat on the reading comprehension exam, they focus on a specific passage since each passage and the associated questions are generally on the same page. Thus, on the reading exam, it is difficult to distinguish between instances of cheating and honest passage knowledge by the students.

[27] Although there is no reason for such classrooms to have elevated test score gains or especially large losses in the following year, which reduces concern that teacher emphasis on specific subject material will lead us to exaggerate the degree of cheating.

in the typical classroom the three most suspicious questions were spread over 2.4 different types of items, and in only 7.7 percent of classrooms did all three questions fall within the same type.

The coefficient estimates in Table 8 suggest that the patterns in classrooms categorized as cheaters were only slightly different than in other classrooms. Focusing on the basic specification in the top row of the table, the point estimates in columns (1) and (3) are small and statistically insignificant. The coefficient in column (4) is significantly higher in the cheating classrooms. That result, however, appears to be an artifact of the particular classification of item types. When one uses a more disaggregated classification scheme (the second row), or excludes the very broad category of numeration (the third row), any evidence that cheating classrooms tend to have more concentration in their suspicious answers disappears. These results are not sensitive to the inclusion of a variety of classroom and school covariates.

VII. Does Teacher Cheating Respond to Incentives?

From the perspective of economics, perhaps the most interesting question related to teacher cheating is the degree to which it is sensitive to incentives. As noted in the introduction, there were two major changes in the incentives faced by teachers and students over our sample period. Prior to 1996, ITBS scores were primarily used to provide teachers and parents with a sense of how a child was progressing academically. Beginning in 1996 with the appointment of Paul Vallas as CEO of Schools, the CPS launched an initiative designed to hold students and teachers accountable for student learning.

The reform had two main elements. The first was putting schools "on probation" if less than 15 percent of students scored at or above national norms on the ITBS reading exams.[28] Probation schools that do not exhibit sufficient improvement may be reconstituted, a procedure that involves closing the school and dismissing or reassigning all of the school staff.[29] It is clear from our discussions with teachers and administrators that being on probation is viewed as an extremely undesirable circumstance. The second piece of the accountability reform was an end to social promotion – the practice of passing students to the next grade regardless of their academic skills or school performance. Under the new policy, students in third, sixth and eighth grade must meet minimum standards on the ITBS in both reading and mathematics in order to be promoted to the next grade. The promotion standards were implemented in Spring 1996 for eighth grade students and in Spring 1997 for third and sixth graders. Promotion decisions are based solely on scores in reading comprehension and mathematics.[30]

Table 9 presents OLS estimates of the relationship between teacher cheating and a variety of classroom and school characteristics.[31] The unit of observation is a classroom*subject*grade*year. The dependent variable is an indicator of whether the classroom cheated. Here we define cheating using our 95th percentile cutoff—that is, a classroom is

[28] The CPS did not use math performance to determine probation status.

[29] Seven high schools have been reconstituted to date, although no elementary schools have suffered this fate. For a more detailed analysis of the probation policy, see Jacob (2001) and Jacob and Lefgren (2001b).

[30] In 1997, the promotion standards for third, sixth and eighth grade were 2.8, 5.3, and 7.0 respectively, which roughly corresponded to the 20th percentile in the national achievement distribution. Students who do not meet the standard in June are required to attend a six-week summer school program, after which they retake the exams. Those students who pass the August exams move on to the next grade. Students who again fail are required to repeat the grade, with the exception of 15-year-olds who attend newly created "transition" centers. In 1997, roughly 30-40 percent of the students in these grades attended summer school and 20 percent of third graders and 12 percent of sixth and eighth graders were retained. For a more detailed analysis of the social promotion policy, see Jacob (2001) and Jacob and Lefgren (2001a).

[31] Logit models evaluated at the mean yield comparable results, so the estimates from a linear probability model are presented for ease of interpretation.

designated a cheater if its SCORE and ANSWERS are above the 95th percentile in that grade, subject and year.[32] In column 1, the policy changes are restricted to have a constant impact across all classrooms. We see that the introduction of the social promotion and probation policies is positively correlated with the likelihood of classroom cheating, although the point estimates are not statistically significant at conventional levels. However, cheating does appear to be responsive to other costs and benefits. Classrooms that tested poorly last year are much more likely to cheat. For example, a classroom with average student prior achievement one classroom standard deviation below the mean is 23 percent more likely to cheat. Classrooms with students in multiple grades are 65 percent less likely to cheat than classrooms where all students are in the same grade. This is consistent with the fact that it is likely more difficult for teachers in such classrooms to cheat, since they must administer two different test forms to students, which will necessarily have different correct answers. Moreover, classes with a higher proportion of students who are included in the official test reporting are more likely to cheat—a 10 percentage point increase in the proportion of students in a class who test scores "count" will increase the likelihood of cheating by roughly 20 percent. Teachers who administer the exam to their own students are 0.67 percentage points—approximately 50 percent—more likely to cheat. Finally, there is no statistically significant impact on cheating of reusing a test form that has been administered in a previous year. That finding is of interest because it suggests that teachers taking old exams and teaching the precise questions to students is not an important component of what we are detecting as cheating (although anecdotal evidence suggests this practice exists).

[32] The results are not sensitive to the cheating cutoff used. Note that this measure may include error due to both false positives and negatives. Since the measurement error is in the dependent variable, it will simply decrease the precision of our estimates.

Much more interesting results emerge when we interact the policy changes with the previous year's test scores for the classroom. For both probation and social promotion, cheating rates in the lowest performing classrooms prove to be quite sensitive to the change in incentives. In column 2, a classroom one-standard deviation below the mean increases cheating by 0.43 percentage points in response to the school probation policy and roughly 0.65 percentage points due to the ending of social promotion. Given the baseline cheating rate of 1.1 percent, these effects are substantial. The magnitude of these changes are particularly large considering that no elementary school on probation has ever been reconstituted since this policy was put into place, and that the social promotion policy has a direct impact on students, but not obvious ramifications for teacher pay or rewards.[33] A classroom one standard deviation above the mean does not see any significant change in cheating in response to these two policies. Such classes are very unlikely to be located in schools at risk for being put on probation, and also are likely to have few students at risk for being retained. The specification shown in column 3 includes a number of classroom and school characteristics, which do not appear to change the coefficients on the policy variables. Consistent with the prior achievement results, classrooms in schools with lower achievement, higher poverty rates and more Black students are more likely to cheat. Interestingly, classrooms in schools with higher quality teachers are less likely to cheat while

[33] While this trend is particularly disturbing due to the relatively minor incentives, given the relatively small number of classrooms engaged in cheating, it is unlikely that such explicit test manipulation has had a large impact on the average achievement levels in Chicago, or the observed increase in achievement since the introduction of high-stakes testing. Table 8 suggests that the accountability policies increased the likelihood of cheating in any one subject by roughly 0.5 to 1.0 percentage points (depending on whether the effects of social promotion and probation are additive and which grade/subject one considers). Now suppose that a teacher manipulates the reading exams for his or her students in a manner that artificially raises their test scores by 2 grade equivalent (which is much larger than we actually observe in the data). This would inflate the system-wide average reading levels by only 0.01 GEs.

classrooms in schools with younger teachers are more likely to cheat. Column 4 illustrates that the effects on the policy variables are robust to including school*year fixed effects.[34]

It is also interesting to examine for which students teachers change answers when they do cheat. One might imagine that teachers would cheat for lower-achieving students, or those who the teacher believed could or should have done better on the exam. However, it is not clear how precisely teachers are able to target their cheating behavior. It is likely that time pressure and concerns about detection will limit the time teachers spend on cheating, and a brief inspection of an answer key will only provide a rough idea of how the student would score without manipulation of his or her answers. Table 10 presents estimates of the relationship between observable student characteristics and cheating. The unit of analysis is a student and the sample is restricted to students in classrooms that were categorized as cheating using the 95th percentile cutoff. The dependent variable takes on the value of one if an individual student's answer string and test score pattern was suspicious at the 95th percentile level, suggesting that the teacher had cheated for that student in the particular subject and year. The results are not sensitive to the particular cutoffs used. The first two columns include all cheating schools; the final two columns narrow the sample to low-achieving schools, where the cheating appears to be concentrated. All of the specifications include fixed effects for classroom*year so that the coefficients are estimated off of variation across students within a particular classroom. Because a student's test score at t-1 is highly correlated with the cheating indicator (by definition), the

[34] Another possible incentive that teachers might respond to is the likelihood of punishment. Punishment for cheating, however, is extremely rare, with only two known instances of cheating teachers being disciplined. Beginning in 1996, CPS began doing audits of test scores. Initially, these were mostly random in nature. More recently, they have focused on classrooms with large test score gains. Our data on audits is incomplete, however. When we included information on audits in the regressions, no statistically significant coefficients were obtained.

equations are estimated using 2SLS where a student's test scores at t-2 are used to instrument for the student's t-1 achievement level.

In column 1, teachers are roughly 6 percentage points more likely to cheat for students who scored in the second quartile (between the 25th and 50th percentile) in the prior year, as compared to students scoring at the third or fourth quartiles. Interestingly, teachers appear least likely to cheat for the lowest-achieving students (the coefficient on the bottom quartile indicator is negative although not statistically significant). Teachers are also less likely to cheat for students who are excluded from test reporting, as would be expected. Teachers also appear to less frequently cheat for boys and for older students.

Column 2 presents an alternative specification that includes a linear measure of the student's prior achievement along with the interaction between prior achievement and an indicator for the high-stakes testing regime (encompassing both the probation and social promotion policies).[35] Here we see that, prior to the introduction of the accountability policy, teachers were more likely to cheat for higher achieving students. The shift by teachers to cheating for lower-achieving students after accountability measures were introduced is consistent with the change in incentives. School probation is based on the fraction of students exceeding a minimum threshold of competence. Student promotion requires a student meeting a hurdle well below the median student in the system. Columns 3 and 4 demonstrate that the same basic relationships hold for the subset of lower-achieving schools.

[35] Ideally, one would also like to include interactions between prior student achievement and the social promotion and school probation policies in specifications that parallel Column 1. Unfortunately, the standard errors become so large that no useful conclusions can be drawn.

VIII. Conclusion

This paper develops an algorithm for determining the prevalence of teacher cheating on standardized tests and applies the results to data from the Chicago Public Schools. Our methods reveal over 1,000 separate instances of classroom cheating, representing 4-5 percent of the classrooms. Moreover, we find that teacher cheating appears quite responsive to relatively minor changes in incentives.

Our results suggest that the implementation of test-based accountability in schools must be approached with caution. If accountability policies create strong incentives without instituting safeguards against cheating, we would predict substantial increases in teacher (or administrator) cheating. This will not only decrease the effectiveness of the reform in identifying struggling schools or highlighting effective pedagogical practices, but will also undermine public confidence in school reform. As high-stakes testing becomes increasingly widespread, these concerns will grow. Fortunately, there are several, relatively inexpensive ways in which schools systems might prevent cheating. Districts could hire an outside agency to proctor the exams rather than having teachers administer the tests. Similarly, teachers in one school might be required to administer exams at another school.

More generally, this paper fits into a small but growing body of research focused on identifying corrupt or illicit behavior on the part of economic actors (e.g. Porter and Zona 1993, Fisman 2000, Di Tella and Schargrodsky 2001, Duggan and Levitt, forthcoming). Because individuals engaged in such behavior actively attempt to cover their trails, the intellectual exercise associated with uncovering their misdeeds differs substantially from the typical economic application in which the researcher starts with a well defined measure of the outcome

variable (e.g. earnings, economic growth, profits) and then attempts to uncover the determinants of these outcomes. In the case of corruption, there is typically no clear outcome variable, making it necessary for the researcher to employ non-standard approaches in generating such a measure. We hope that the methods utilized in this paper provide some guidance to those seeking to identify corruption in other domains.

References

Aiken, L.R. (1991). Detecting, understanding and controlling for cheating on tests. *Research in Higher Education*, 32(6), 725-736.

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69(345), 44-49.

Bellezza, F.S. and Bellezza, S.F. (1989). Detection of Cheating on Multiple-Choice Tests by Using Error Similarity Analysis. *Teaching of Psychology*, 16(3), 151-155.

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.

DiTella, Rafael, and Ernesto Schargrodsky. (2000). "The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires," Unpublished manuscript, Harvard Business School.

Duggan, Mark, and Steven Levitt. (Forthcoming). "Winning Isn't Everything: Corruption in Sumo Wrestling," *American Economic Review*.

Fisman, Ray, (Forthcoming). "Estimating the Value of Political Connections," *American Economic Review*.

Frary, R.B. (1993). Statistical detetection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education*, 6(2), 153-165.

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.

Holland, P.W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Technical Report No. 96-5). Princeton, NJ: Educational Testing Service.

Gay, G. H. (1990). "Standardized Tests: Irregularities in Administering of Tests Affect Test Results," *Journal of Instructional Psychology* 17(2):93-103.

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

Hofkins, D. (1995, June 16). Cheating "rife" in national tests. Times Educational Supplement, p. 1.

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

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

Jacob, B. (2001b). The Impact of Test-Based Accountability in Schools: Evidence from Chicago. Unpublished manuscript. John F. Kennedy School of Government, Harvard University.

Jacob, B. and Lefgren, L. (2001a). The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago. Unpublished manuscript. John F. Kennedy School of Government, Harvard University.

Jacob, B. and Lefgren, L. (2001b). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. Unpublished manuscript. John F. Kennedy School of Government, Harvard University.

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.

Ladd, H. F. (1999). "The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes." Economics of Education Review **18**: 1-16.

Lindsay, D. (1996, October 2). Whodunit? Officials find thousands of erasures on standardized tests and suspect tampering. *Education Week*, 25-29.

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.

Perlman, C.L. (1985, March). *Results of a Citywide Testing Program Audit in Chicago.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC ED 263 212).

Porter, Robert H. and J. Douglas Zona, "Detection of Bid Rigging in Procurement Auctions," *Journal of Political Economy*, CI (1993), 518-538.

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.

Shepard, L.A. and Doughtery, K.C. (1991). *Effects of High-Stakes testing on Instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC ED 337 468).

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, R. L. (2001). The Influence of High-Stakes Testing on Instructional Practice in Chicago. American Educational Research Association, Seattle, WA.

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

Wollack, J.A. (1997). A Nominal Response Model Approach for Detecting Answer Copying. *Applied Psychological Measurement*, 22(2), 144-152.

Appendix A: The Construction of Suspicious String Measures

We rely on two different indicators of cheating: (1) unusually large test score gains that are not sustained on future exams and (2) unexpected response patterns among students within the same classroom. We measure the likelihood of classroom response patterns in four different ways. This appendix describes in greater detail how we construct each of the four measures of unexpected or suspicious responses.

The first measure focuses on the most unlikely block of identical answers given on consecutive questions. This is meant to pick up teachers who change a series of questions for some number of students in their classroom. For example, a teacher may fill in the correct responses for the last six questions on the exam for ten low-achieving students in the class. We calculate the probability that this block of answers would have occurred if student responses within a classroom were uncorrelated. The more unlikely is the most unexpected block of test responses, the more likely it is that cheating occurred.

Using past test scores, future test scores and background characteristics, we predict the likelihood that each student will give each answer on each question. For each item, a student has four choices (A, B, C or D), only one of which is correct. We estimate a multinomial logit for each item on the exam in order to predict how students will respond to each question. We estimate the following model for each item, using information from other students in that year, grade and subject.

$$(1) \qquad \Pr(Y_{isc} = j) = \frac{e^{\beta_j x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}}$$

where Y_{isc} indicates the response of student s in class c on item i, the number of possible responses (J) is four, and and X_s is a vector that includes measures of prior and future student achievement in math and reading as well as demographic variables (such as race, gender and free lunch status) for student s. Thus, a student's predicted probability of choosing a particular response is identified by the likelihood of other students (in the same year, grade and subject) with similar background characteristics choosing that response.

Notice that by including future as well as prior test scores in the model we decrease the likelihood that students with unusually good teachers will be identified as cheaters, since these students will likely retain some of the knowledge learned in the base year and thus have higher future test scores. Also note that by estimating the probability of selecting each possible response, rather than simply estimating the probability of choosing the correct response, we take advantage of any additional information that is provided by particular response patterns in a classroom.

Using the estimates from this model, we calculate the predicted probability that each student would answer each item in the way that he or she in fact did.

(2) $$p_{isc} = \frac{e^{\hat{\beta}_k X_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j X_s}}$$ for k = response actually chosen by student s on item i

This provides us with one measure per student per item. Taking the product over items within student, we calculate the probability that a student would have answered a string of consecutive questions from item m to item n as he or she did:

(3) $$p_{sc}^{mn} = \prod_{i=m}^{n} \hat{p}_{isc}$$

We then take the product across all students in the classroom who had identical responses in the string. If we define z as a student, S_{zc}^{mn} as the string of responses for student *z* from item *m* to item *n*, and $\bar{S}_{sc}^{mn}$ and as the string for student *s*, then we can express the product as:

$$(4) \qquad \tilde{p}_{sc}^{mn} = \prod_{s \in \{z : S_{zc}^{mn} = \bar{S}_{sc}^{mn}\}} \hat{p}_{sc}^{mn}$$

Note that if there are *ns* students in class *c*, and each student has a unique set of responses to these particular items, then $\tilde{p}_{sc}^{mn}$ collapses to p_{sc}^{mn} for each student and there will be *ns* distinct values within the class. On the other extreme, if all of the students in class *c* have identical responses, then there is only one distinct value of $\tilde{p}_{sc}^{mn}$. We repeat this calculation for all possible consecutive strings of length three to seven; that is for all S^{mn} such that $3 \le m - n \le 7$. We have experimented with searching over longer strings, but this does not change our results.

To create our first indicator of suspicious string patterns, we take the minimum of the predicted block probability for each classroom.

Measure 1: $M1_c = \min_s(\tilde{p}_{sc}^{mn})$

This measure captures the least likely block of identical answers given on consecutive questions in the classroom.

The second measure of suspicious answer strings is intended to capture more general patterns of similarity in student responses. When a teacher changes answers on student test forms, it presumably increases the uniformity of responses across students in the class. Thus, the overall degree of correlation in student answers across the test may be quite high, even if there is not one particularly unusual block of identical answers.

To construct this measure, we first calculate the residuals for each of the possible choices a student could have made for each item.

$$e_{jisc} = 0 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{ if } j \neq k$$

(5)

$$= 1 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{ if } j = k$$

where e_{jisc} is the residual for response j on item i by student s in classroom c. We thus have four separate residuals per student per item.

To create a classroom level measure of the response to item i, we need to combine the information for each student. First, we sum the residuals for each response across students within a classroom.

$$e_{jic} = \sum_{s} e_{jisc} \quad (6)$$

If there is no within class correlation in the way that students responded to a particular item, this term should be approximately zero. Second, we sum across the four possible responses for each item within classrooms. At the same time, we square each of the component residual measures to accentuate outliers and divide by number of students in the class (ns_c) to normalize by class size.

$$v_{ic} = \frac{\sum_{j} e_{jic}^2}{ns_c} \quad (7)$$

The statistic v_{ic} captures something like the variance of student responses on item i within classroom c. Notice that we choose to first sum across the residuals of each response across

students and then sum the classroom level measures for each response, rather than summing across responses within student initially. We do this in order to emphasize the classroom level tendencies in response patterns.

Our second measure of suspicious strings is simply the classroom average (across items) of this variance term across all test items.

Measure 2: $M2_c = \bar{v}_c = \frac{\sum_i v_{ic}}{ni}$ where *ni* is the number of items on the exam.

Note that within-classroom correlation may arise for many reasons other than cheating. For example, a teacher may emphasize a certain topic or set of skills during the school year.

Our third measure focuses on the *variance* (as opposed to the mean) in the degree of correlation across questions. If the teacher changes answers for multiple students on some set of questions, the within-classroom correlation on those particular items will be extremely high while the degree of within-classroom correlation on other questions will likely be typical. This will cause the cross-question variance in correlations to be larger than normal in cheating classrooms.

Measure 3: $M3_c = \sigma^2_{v_c} = \frac{\sum_i (v_{ic} - \bar{v}_c)^2}{ni}$

Our final indicator focuses on the extent to which a student's response pattern was different from other student's with the same aggregate score that year. Questions vary significantly by difficulty. The typical student will answer most of the easy questions correctly and get most of the hard questions wrong. If students in a class miss the easy questions while answering the hard questions correctly, this could be an indicator of cheating.

Let q_{isc} equal one if student s in classroom c answered item i correctly, and zero otherwise. Let A_s equal the aggregate score of student s on the exam. We then determine what fraction of students at each aggregate score level answered each item correctly. If we let ns_A equal then number of students with an aggregate score of A, then this fraction, $\overline{q}_i^A$, can be expressed as

$$(8) \qquad \overline{q}_i^A = \frac{\sum_{s \in \{r \mid A_r = A_s\}} q_{isc}}{ns_A}$$

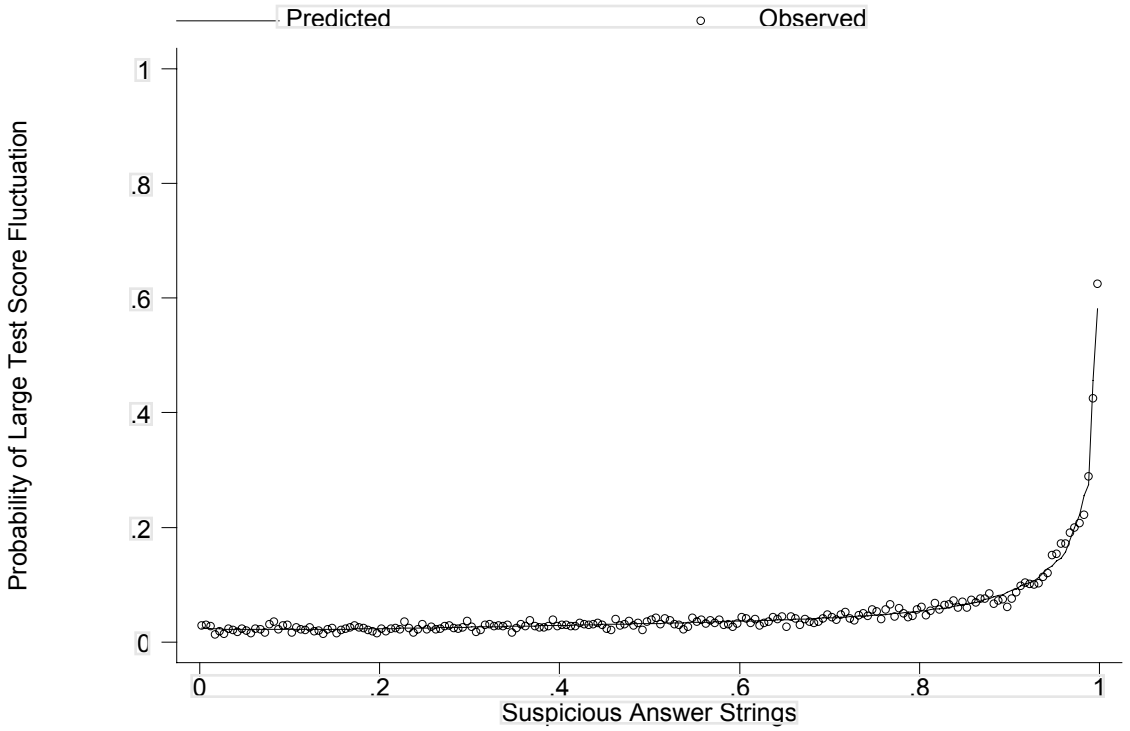
We then calculate a measure of how much the response pattern of student s differed from the response pattern of other students with the same aggregate score. We do so by subtracting a student's answer on item i from the mean response of all students with aggregate score A, squaring these deviations and then summing across all items on the exam.

$$(9) \qquad Z_{sc} = \sum_i \left(q_{isc} - \overline{q}_i^A\right)^2$$

We then subtract out the mean deviation for all students with the same aggregate score, $\overline{Z}^A$, and sum the students within each classroom to obtain our final indicator.

Measure 4: $M4_c = \sum_s \left(Z_{sc} - \overline{Z}^A\right)$

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.

Figure 2: Sample Answer Strings and Test Scores from Two Classrooms

Student Answer Strings (each row represents one student's answers)	Student Test Scores		
	Year t-1	**Year t**	**Year t+1**
Suspected Cheating Classroom			
112A4A342CB214D0001ACD24A3A12DADBCB4A0000000	1.9	5.3	4.4
1B2A34D4AC42D23B141ACD24A3A12DADBCB4A2134141	4.3	5.6	4.3
DB2ABAD1ACBDDA212B1ACD24A3A12DADBCB400000000	3.0	6.5	5.1
D43A3A24ACB1D32B412ACD24A3A12DADBCB422143BC0	5.2	5.9	4.9
D43AB4D1AC3DD43421240D24A3A12DADBCB400000000	4.8	5.3	3.6
1142340C2CBDDADB4B1ACD24A3A12DADBCB43D133BC4	3.6	6.3	4.9
DBA2BA21AC3D2AD3C4C4CD40A3A12DADBCB400000000	1.9	6.1	3.6
DBAA4ADC4CBD24DBCB2A1110A3A12DADBCB400000000	3.3	6.3	6.2
144A3ADC4CBDDADBCBC2C2CC43A12DADBCB4211AB343	3.0	6.8	4.9
D43ABA3CACBDDADBCBCA42C2A3212DADBCB42344B3CB	4.8	7.1	6.6
214AB4DC4CBDD31B1B2213C4AD412DADBCB4ADB00000	3.6	6.1	4.3
313A3AD1AC3D2A23431223C000012DADBCB400000000	3.8	4.7	5.1
D4AAB2124CBDDADBCB1A42CCA3412DADBCB423134BC1	5.5	6.6	7.7
3B3AB4D14C3D2AD4CBCAC1C003A12DADBCB4ADB40000	3.0	6.5	6.6
DBAAB3DCACB1DADBC42AC2CC31012DADBCB4ADB40000	3.8	7.1	5.6
DB223A24ACB11A3B24CACD12A241CDADBCB4ADB4B300	4.9	6.5	5.8
D122BA2CACBD1A13211A2D02A2412D0DBCB4ADB4B3C0	3.6	6.1	6.2
1423B4D4A23D24131413234123A243A2413A21441343	4.9	2.5	5.6
DB4ABADCACB1DAD3141AC212A3A1C3A144BA2DB41B43	5.9	6.5	7.7
DB2A33DCACBD32D313C21142323CC300000000000000	3.8	4.4	5.6
1B33B4D4A2B1DADBC3CA22C0000000000000000000	5.0	4.4	7.2
D12443D43232D32323C213C22D2C23234C332DB4B300	3.3	3.8	3.6
D4A2341CACBDDAD3142A2344A2AC23421C00ADB4B3CB	6.4	5.9	6.2
* * ** **!*!!!!!!!!!*! ! *	**4.1**	**5.8**	**5.5**
Correlation across students on each question (* = high , ! = very high)	Average Test Scores		
Typical Classroom			
DB3A431422BD131B4413CD4221A1CDA332342D3AB4C4	4.0	5.1	5.1
D1AA1A11ACB2D3DBC1CA22C23242C3A142B3ADB243C1	4.6	5.9	5.3
D42A12D2A4B1D32B21CA2312A3411D000000000000000	4.5	3.8	6.4
3B2A34344C32D21B1123CDC0000000000000000000000	3.3	2.8	5.1
34AABAD12CBDD3D4C1CA112CAD2CCD000000000000000	3.8	5.6	6.4
D33A3431A2B2D2D44B2ACD2CAD2C2223B40000000000	4.6	4.9	5.8
23AA32D2A1BD2431141342C13D212D233C34A3B3B000	3.3	4.4	4.9
D32234D4A1BDD23B242A22C2A1A1CDA2B1BAA33A0000	5.1	5.6	5.9
D3AAB23C4CBDDADB23C322C2A222223232B443B24BC3	4.7	5.6	7.0
D13A14313C31D42B14C421C42332CD2242B3433A3343	2.2	3.8	4.9
D13A3AD122B1DA2B11242DC1A3A12100000000000000	4.5	4.1	5.9
D12A3AD1A13D23D3CB2A21CCADA24D2131B440000000	3.6	5.3	5.9
314A133C4CBD142141CA424CAD34C122413223BA4B40	3.3	4.7	4.4
D42A3ADCACBDDADBC42AC2C2ADA2CDA341BAA3B24321	5.6	6.9	8.5
DBAA34DC2CB2DADB24C412C1ADA2C3A341BA20000000	5.0	5.9	7.0
D1341431ACBDDAD3C4C213412DA22D3D1132A1344B1B	3.8	5.3	5.3
1BA41A21A1B2DADB24CA22C1ADA2CD32413200000000	4.3	5.3	6.8
DBAA33D2A2BDDADBCBCA11C2A2ACCDA1B2BA20000000	4.5	6.8	7.9
* * ** *	**4.2**	**5.1**	**6.0**

Notes: The data in the table represent actual answer strings and test scores from two CPS classrooms taking the same exam. The top classroom is suspected of cheating; the bottom classroom is not. Each row corresponds to an individual student. Each column represents a particular question on the exam. A letter indicates that the student gave that answer and the answer was correct. A number means that the student gave the corresponding letter answer (e.g. 1="A"), but the answer was incorrect. A value of "0" means the question was left blank. Student test scores, in grade equivalents, are shown in the last three columns of the table. The test year for which the answer strings are presented is denoted year t. The scores from years t-1 and t+1 correspond to the preceding and following years' examinations.

Table 1: Summary Statistics

Variables	Cross-Classroom Analysis (classroom level data, one observation per class*year*subject	Within-Classroom Analysis (student level data, includes only students in classrooms labeled cheaters using the 95th % cutoff)
Accountability Policy		
Social promotion policy	0.215	0.289
School probation policy	0.127	0.164
Test form offered for the first time	0.371	0.327
Student Characteristics		
In bottom quartile of achievement in prior year	--	0.416
In second quartile of achievement in prior year	--	0.283
In third quartile of achievement in prior year	--	0.202
Excluded from official test reporting in current year	--	0.041
Reading percentile 2 years prior	--	34.3
Math percentile 2 years prior	--	38.3
Black	--	0.732
Hispanic	--	0.194
Male	--	0.486
Age	--	11.0 (1.5)
Special Education	--	0.061
Living in foster care	--	0.050
Living with non-parental relative	--	0.076
Teacher cheated for this student – 80th percentile cutoff	--	0.448
Teacher cheated for this student – 90th percentile cutoff	--	0.294
Teacher cheated for this student – 95th percentile cutoff	--	0.180
Classroom Characteristics		
Mixed grade classroom	0.073	0.021
Teacher administers exams to her own students (3rd grade)	0.206	0.291
Percent of students who were tested and included in official reporting	0.883	0.915
Average prior achievement (as deviation from year*grade*subject mean)	-0.004 (0.661)	-0.151 (0.558)
% Black	0.595	0.726
% Hispanic	0.263	0.196
% Male	0.495	0.491
% Old for grade	0.086	0.082
% Living in foster care	0.044	0.053
% Living with non-parental relative	0.104	0.082
Cheater – 95th percentile cutoff	0.013	1.00
School-Level Teacher Characteristics		
Average quality of teachers' undergraduate institution in the school	-2.550 (0.877)	-2.801 (0.846)
Percent of teachers who live in Chicago	0.712	0.723
Percent of teachers who have a MA or PhD	0.475	0.480
Percent of teachers who majored in education	0.712	0.719
Percent teachers under 30 years of age	0.114	0.111
Percent of teachers at the school less than 3 years	0.547	0.546
School Characteristics		
% students at national norms in reading last year	28.8	26.0

% students receiving free lunch in school	84.7	88.5
Predominantly Black school	0.522	0.687
Predominantly Hispanic school	0.205	0.164
Mobility rate in school	28.6	29.5
Attendance rate in school	92.6	92.3
School size	722 (317)	784 (304)
Number of observations	163,474	39,216

Notes: Robust standard errors clustered by school*year are shown in parenthesis. Other variables included in the regressions but not shown here include cubic terms for the number of students in the class as well as indicators of the percent of students who were black, Hispanic, receiving free lunch, old for grade, in a special education program, male, living in foster care, and living with a non-parental relative.

Table 2: The Fraction of Classrooms Identified as Cheaters Using Simulated Data for 5th Grade Math in 1993

		Percent of class affected		
Manner in which the test forms are altered	Number of questions changed per student	25	50	100
Scenario I – Unsophisticated cheater Teacher changes a randomly selected block of answers for a randomly selected group of students. Students do not retain any gains.	3	3.95	20.06	56.40
	6	15.16	53.95	87.29
	9	29.47	77.68	89.45
Scenario II – Sophisticated cheater Teacher changes every other incorrect question for randomly selected students in the class. Students do not retain any gains.	3	2.26	10.08	52.54
	6	5.56	33.80	86.06
	9	11.68	57.25	91.43
Scenario III – A Good teacher who provides real gains for students Teacher enhances learning so that students correctly answer marginal questions. Students retain 80% of gains.	3	0.75	0.75	2.64
	6	0.75	2.64	17.04
	9	1.41	7.91	37.48

Notes: The results in this table are from simulations in which the authors alter test answers in an attempt to imitate cheating or outstanding teaching. For each classroom, we manipulate the answer strings in the manner stated in the table, and then determine whether the classroom would qualify as cheating by our definition, holding constant the other classrooms in that grade and year. The results presented are for fifth grade reading in 1993, using our measure of cheating based on the 95th percentile of both ANSWERS and SCORE. The baseline cheating rate in the raw data for this subject, grade and year is 1.13 percent.

Table 3: The Relationship between Measures of Unusual Test Scores and Suspicious Answer Strings in Parts of the Distribution Unlikely to Contain Many Cheating Classrooms

		ANSWERS falls within the range		
		0-25th percentile	25-50th percentile	50-75th percentile
Percent of observations with SCORE above:	80th percentile	0.163	0.175	0.187
	90th percentile	0.062	0.073	0.088
	95th percentile	0.023	0.028	0.038
		SCORE falls within the range		
		0-25th percentile	25-50th percentile	50-75th percentile
Percent of observations with ANSWERS above:	80th percentile	.245	.151	.137
	90th percentile	.118	.062	.060
	95th percentile	.049	.026	.026

Notes: Values in the table are the percentage of classrooms in the sample meeting the criteria of each cell in a particular year on a particular subject test. The unit of observation is a classroom*subject*year. If SCORE and ANSWERS were independently distributed, the values in the first and fourth rows of the table will be .20, in the second and fifth rows the values will be .10, and in the 3rd and sixth rows the values will be .05.

Table 4: Estimated Prevalence of Teacher Cheating
(Percent of All Classrooms in a Single Year)

Percent cheating on a particular test subject (e.g.Reading comprehension, Math 1)				
		Cutoff for Test Score Fluctuations (SCORE):		
Cutoff for suspicious answer strings (ANSWERS)		80th percentile	90th percentile	95th percentile
	80th percentile	2.1	2.1	1.8
	90th percentile	1.8	1.8	1.5
	95th percentile	1.3	1.3	1.1
Percent cheating on at least one of the four tests given				
		Cutoff for Test Score Fluctuations (SCORE):		
Cutoff for suspicious answer strings (ANSWERS)		80th percentile	90th percentile	95th percentile
	80th percentile	4.5	5.6	5.3
	90th percentile	4.2	4.9	4.4
	95th percentile	3.5	3.8	3.4

Notes: The top panel of the table presents estimates of the percentage of classrooms cheating on a particular subject test in a given year based on three alternative cutoffs for *ANSWERS* and *SCORE*. In all cases, the prevalence of cheating is based on the excess number of classrooms with unexpected test score fluctuation among classes with suspicious answer strings relative to classes that do not have suspicious answer strings. The bottom panel of the table presents estimates of the percentage of classrooms cheating on at least one of the four subject tests that comprise the overall test. In the bottom panel, classrooms that cheat on more than one subject test are only counted once. Our sample includes over 35,000 3rd-7th grade classrooms in the Chicago Public Schools for the years 1993-1999.

Table 5: The Prevalence of Cheating by Grade and Subject

	95th Percentile			
Grade	Reading Comprehension	Math 1 (Number Concepts & Estimation)	Math 2 (Data Interpretation & Problem Solving)	Math 3 (Computation)
3rd	2.41	1.46	1.41	1.23
4th	1.29	1.26	0.76	1.04
5th	0.90	1.04	0.49	1.30
6th	1.18	1.05	0.78	1.42
7th	0.63	0.80	0.32	0.95
	90th Percentile			
Grade	Reading Comprehension	Math 1 (Number Concepts & Estimation)	Math 2 (Data Interpretation & Problem Solving)	Math 3 (Computation)
3rd	4.05	2.72	2.50	1.92
4th	1.72	1.90	1.19	1.64
5th	1.04	1.47	0.64	2.09
6th	1.39	1.93	0.89	2.55
7th	1.10	1.49	0.74	1.64

Notes: The top panel of the table presents estimates of the percentage of classrooms cheating on a particular subject test in a given year based on the 90th percentile cutoff for *ANSWERS* and *SCORE* and the bottom panel presents those for the 90th percentile cutoff. In all cases, the prevalence of cheating is based on the excess number of classrooms with unexpected test score fluctuation among classes with suspicious answer strings relative to classes that do not have suspicious answer strings.

Table 6: Mean Reversion in Classrooms with Large Test Score Gains Categorize by Suspiciousness of Answer Strings

	Class percentile rank on suspiciousness of answer strings				
	Low (0-50th)	Moderate (50-80th)	High (80-95th)	Very High (95-99th)	Highest (>99th)
Panel A: Sample includes top 10% of classrooms on average test score gain measure					
Current year test score gain (relative to the system mean)	0.59	0.61	0.67	0.78	0.93
Subsequent year test score gain (relative to the system mean)	-0.11	-0.21	-0.34	-0.53	-0.81
Percent of excess gain lost in the following year	81.0	65.6	49.3	32.1	12.9
Number of classrooms	6,781	4,190	3,094	1,616	884
Panel B: Sample includes top 5% of classrooms on average test score gain measure					
Current year test score gain (relative to the system mean)	0.71	0.74	0.79	0.91	1.01
Subsequent year test score gain (relative to the system mean)	-0.13	-0.25	-0.38	-0.60	-0.86
Percent of excess gain lost in the following year	81.7	66.2	51.9	34.1	14.9
Number of classrooms	2,858	1,964	1,674	1,054	735
Panel C: Sample includes top 1% of classrooms on average test score gain measure					
Current year test score gain (relative to the system mean)	0.99	1.00	1.07	1.20	1.26
Subsequent year test score gain (relative to the system mean)	-0.19	-0.29	-0.47	-0.80	-1.06
Percent of excess gain lost in the following year	80.8	71.0	56.1	33.3	15.9
Number of classrooms	332	288	335	353	356

Notes: Values reported in the top two rows of the table are the excess test score gains in the current year and the following year for the ten percent of classrooms experiencing the greatest test score gains, broken down by how suspicious the classes answer strings are in the current year. Excess test scores are defined as the mean test score gain in the class (measured in grade equivalents) relative to the system mean in that grade, subject, and year. The third row of the table presents the fraction of the excess gain that a classroom loses in the following year. The larger is this number, the more transitory were the previous year's gains. Spurious gains due to cheating are expected to be more transitory than gains due to true learning.

Table 7: Patterns of Cheating within Classrooms and Schools

	Dependent variable = Class suspected of cheating (Class is above the 95th percentile on both SCORE and ANSWERS on a particular subject test: mean=0.011)			
Independent Variables	Full Sample		Sample of classes and school that existed in the prior year	
Classroom cheated on exactly one other subject this year on this	0.105	0.103	0.101	0.101
	(0.008)	(0.008)	(0.009)	(0.009)
Classroom cheated on exactly two other subjects this year	0.289	0.285	0.243	0.243
	(0.027)	(0.027)	(0.031)	(0.031)
Classroom cheated on all three other subjects this year	0.627	0.622	0.595	0.595
	(0.051)	(0.051)	(0.054)	(0.054)
Cheating rate among all other classes in the school this year on this subject	--	0.166	0.134	0.129
		(0.030)	(0.027)	(0.027)
Cheating rate among all other classes in the school this year on other subjects	--	0.023	0.059	0.045
		(0.024)	(0.026)	(0.029)
Cheating in this classroom in this subject last year	--	--	0.096	0.091
			(0.012)	(0.012)
Number of other subjects this classroom cheated on last year	--	--	0.023	0.018
			(0.004)	(0.004)
Cheating in this classroom ever in the past	--	--	--	0.006
				(0.002)
Cheating rate among other classrooms in this school in past years	--	--	--	0.090
				(0.040)
Full set of grade*subject*year interactions included?	Yes	Yes	Yes	Yes
R-squared	0.090	0.093	0.109	0.109
Number of Observations	165,578	165,578	94,182	94,170

Notes: The dependent variable is an indicator for whether a classroom is above the stated cutoff on ANSWERS and SCORE on a particular subject test. Estimation is done using a linear probability model. Columns that include measures of cheating in prior years, observations where that classroom and/or school does not appear in the data in the prior year are excluded. Standard errors are clustered at the school level to take into account correlations across classroom as well as serial correlation.

Table 8: The Relationship between within-class Correlations, Item Types, and Cheating

	Dependent variable			
	For the 3 most highly correlated questions …		For the 5 most highly correlated questions …	
Independent Variable = Classroom labeled a cheater	The number of different item types	Whether all of the questions were of a single type	The number of different item types	Whether all of the questions were of a single type
Classroom labeled a cheater	-0.009 (0.027) [2.403]	0.018 (0.011) [0.076]	0.039 (0.037) [3.335]	0.022 (0.005) [0.014]
Alternative Specifications				
Consider estimation sub-groups as separate item categories	-0.068 (0.022) [2.642]	0.020 (0.006) [0.020]	-0.120 (0.033) [3.927]	0.001 (0.001) [0.001]
Exclude numeration items because this category encompasses a wide variety of different math skills (e.g., operations with positive and negative numbers, inequalities, exercises involving number lines, definitions of integer, whole number, fraction, etc.)	0.039 (0.029) [1.983]	-0.020 (0.018) [0.235]	0.122 (0.036) [2.723]	-0.004 (0.011) [0.066]

Notes: The unit of observation for this analysis is classroom*grade*year and the sample is limited to results from the Math Section I exam. The cheating indicator used is based on the 95th percentile cutoff. All estimates include fixed effects for grade and year. Robust standard errors that account for the correlation within a school*year are shown in parenthesis and the baseline (non-cheating) means are shown in square brackets.

Table 9: OLS Estimates of the Relationship between Cheating and Classroom Characteristics

	Dependent variable = Indicator of classroom cheating			
Independent variables	(1)	(2)	(3)	(4)
Social promotion policy	0.0011	0.0011	0.0015	0.0023
	(0.0013)	(0.0013)	(0.0013)	(0.0009)
School probation policy	0.0020	0.0019	0.0021	0.0029
	(0.0014)	(0.0014)	(0.0014)	(0.0013)
Prior classroom achievement	-0.0047	-0.0028	-0.0016	-0.0028
	(0.0005)	(0.0005)	(0.0007)	(0.0007)
Social promotion*classroom achievement	--	0.0049	0.0051	0.0046
		(0.0014)	(0.0014)	(0.0012)
School probation*classroom achievement	--	-0.0070	-0.0070	-0.0064
		(0.0013)	(0.0013)	(0.0013)
Mixed grade classroom	0.0084	0.0085	0.0089	0.0089
	(0.0007)	(0.0007)	(0.0008)	(0.0012)
% of students included in official reporting	0.0252	0.0249	0.0141	0.0131
	(0.0031)	(0.0031)	(0.0037)	(0.0037)
Teacher administers exam to own students	0.0067	0.0067	0.0066	0.0061
	(0.0015)	(0.0015)	(0.0015)	(0.0011)
Test form offered for the first time	-0.0007	-0.0007	-0.0011	--[a]
	(0.0011)	(0.0011)	(0.0010)	
Average quality of teachers' undergraduate institution	--	--	-0.0026	--
			(0.0007)	
Percent of teachers who have worked at the school less than 3 years	--	--	-0.0045	--
			(0.0031)	
Percent teachers under 30 years of age	--	--	0.0156	--
			(0.0065)	
Percent of students in the school meeting national norms in reading last year			0.0001	
			(0.0000)	
Percent free lunch in school	--	--	0.0001	--
			(0.0000)	
Predominantly Black school	--	--	0.0068	--
			(0.0019)	
Predominantly Hispanic school	--	--	-0.0009	--
			(0.0016)	
School*Year Fixed Effects	No	No	No	Yes
Number of observations	163,474	163,474	163,474	163,474

Notes: The unit of observation is classroom*grade*year*subject and the sample includes years eight years (1993 to 2000), four subjects (reading comprehension and three math sections) and five grades (three to seven). The dependent variable is the cheating indicator derived using the 95th percentile cutoff. Robust standard errors clustered by school*year are shown in parenthesis. Other variables included in the regressions in column 1 and 2 include a linear time trend, grade, cubic terms for the number of students, a linear grade variable, and fixed effects for subjects. The regression shown in column 3 also includes the following variables: indicators of the percent of students in the classroom who were black, Hispanic, male, receiving free lunch, old for grade, in a special education program, living in foster care and living with a non-parental relative, indicators of school size, mobility rate and attendance rate, and indicators of the percent of teachers in the school Other variables include the percent of teachers in the school who had a masters or doctoral degree, lived in Chicago and were education majors. [a] Test forms vary only by year so this variable will drop out of the analysis when school*year fixed effects are included.

Table 10: In Cheating Classrooms, for Whom do Teachers Cheat?

	Dependent variable = Teacher cheated for the student			
Independent variables	(1)	(2)	(3)	(4)
Prior achievement in the bottom quartile	0.011 (0.038)	--	-0.007 (0.075)	--
Prior achievement in the 2nd quartile	0.057 (0.024)	--	0.069 (0.039)	--
Prior achievement in the 3rd quartile	0.023 (0.067)	--	-0.012 (0.141)	--
Prior achievement (linear measure)	--	0.0004 (0.0003)	--	0.0005 (0.0004)
Prior achievement (linear) * High-stakes	--	-0.0007 (0.0004)	--	-0.0007 (0.0005)
Excluded from test reporting	-0.045 (0.014)	-0.048 (0.014)	-0.045 (0.021)	-0.052 (0.020)
Male	-0.009 (0.004)	-0.009 (0.004)	-0.014 (0.005)	-0.013 (0.005)
Black	0.005 (0.011)	0.006 (0.011)	0.004 (0.024)	0.001 (0.023)
Hispanic	-0.010 (0.010)	-0.008 (0.009)	0.006 (0.023)	0.004 (0.022)
Age	-0.010 (0.004)	-0.012 (0.004)	-0.015 (0.005)	-0.017 (0.005)
Sample	Full		Low-Achieving Schools	
Number of observations	39,216		23,010	

Notes: The sample includes only those classrooms that were categorized as cheating based on the 95th percentile cutoff in a particular subject and year. The dependent variable takes on the value of one if a *student's* answer string and test score pattern was suspicious at the 90th percentile level, suggesting that the teacher had cheated for that student in the particular subject and year. All models include fixed effects for classroom*year. Low achieving schools are defined as those in which fewer than 25% of students met national norms in reading in 1995. The equations are estimated using 2SLS where a student's test scores at t-2 are used to instrument for the student's t-1 achievement level. Robust standard errors are shown in parenthesis.