

Very preliminary and incomplete
Comments appreciated

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

Brian A. Jacob
Kennedy School of Government
brian_jacob@harvard.edu

Steven D. Levitt
University of Chicago and American Bar Foundation
slevitt@midway.uchicago.edu

Current draft: October 2001

*Preliminary and incomplete. We would like to thank Suzanne Cooper, Mark Duggan, Sue Dynarski, Arne Duncan, and Michael Greenstone for helpful comments and discussions. Financial support was provided by the National Science Foundation and the Sloan Foundation. Addresses: Brian Jacob, Kennedy School of Government, Harvard University, 79 JFK Street, Cambridge, MA 02138; Steven Levitt, Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637.

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 4-5 percent of elementary school classrooms annually. The observed frequency of cheating appears to respond strongly to relatively minor changes in incentives. Our results suggest that introducing high-stakes testing without appropriate safeguards would be likely to lead to widespread cheating.

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 (ECS). 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.

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 (CITE). By linking teacher salary and employment to student test scores, schools are held accountable for their students' performance. Opponents of testing, on the other hand, argue that linking incentives to performance on standardized tests may lead teachers to substitute away from other worthwhile objectives such as teaching skills or subjects not directly tested on the required exam or pursuing broader goals such as teaching civic responsibility (Holmstrom and Milgrom 1991). Indeed, there is evidence that teachers tailor lesson plans to the specific material covered on standardized tests (Heubert and Hauser 1999).

In this paper, we explore a very different concern regarding high-stakes testing—cheating on the part of teachers and administrators.¹ As incentives for high test scores

¹ Hereafter, we use the phrase “teacher cheating” to encompass cheating done by either teachers or administrators.

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 New York (CITE), Texas (CITE), California (CITE), Illinois (CITE), Massachusetts (CITE), and Great Britain (CITE).

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 Dougherty 1991). In another study, 35 percent of North Carolina 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.² 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. In the mid-eighties, Perlman (1985) investigated suspected cheating in a number of Chicago public schools. 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).

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

² 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).

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).³ 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 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).

Not every classroom with test score fluctuations and suspicious answer strings, however, is cheating. Sometimes such patterns arise by chance. The key assumption for our identification strategy is that among classrooms that are not cheating, test score fluctuations and unusual answer strings are uncorrelated (or alternatively, we know the extent of the correlation). This assumption is testable and is supported by the data. Using this assumption and the simple

³ We do not, however, have access to the actual test forms that students filled out so we are unable to analyze these

statistical model we develop, it is straightforward to estimate the number of cheating classrooms. By varying the assumed correlation between our measures in non-cheating classrooms, or the thresholds for what qualifies as an unusual test score fluctuation or a suspicious answer string, we are able to test the sensitivity of our results.

Figure 1 provides visual evidence in support of the empirical approach we take. The horizontal axis in the figure ranks classrooms according to how suspicious their answer strings are according to our measures.⁴ 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.⁵ 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 cheating, we will essentially compare the area under the curve in the right tail of Figure 1 to the predicted area under the curve under our maintained assumptions about how the two measures co-vary in non-cheating classrooms.

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

tests for evidence of suspicious patterns of erasures.

⁴ We defer a precise discussion of how we construct our cheating indicators until Section III.

⁵ The graph plots the fitted values of a regression with a seventh-order polynomial in the measure of suspicious answer strings.

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 classroom, 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 is a strong predictor of cheating on other sections of the test. 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 have lost 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, the most suspicious answer strings for in cheating classrooms do not fall disproportionately within one reading passage, as one might expect if the pattern were driven by a teacher who happened to cover a particular story or topic during the school year. Finally, the most unexpected math responses within cheating classrooms do not cluster within a single topic (e.g., all algebra items, or all questions about fractions) as one might expect if the teacher had simply emphasized certain material during the course.

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 when Paul Vallas assumed leadership of the CPS 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 in 1996-97, 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 provides a brief overview of the institutional details of the Chicago Public Schools and the data set that we use. Section IV introduces the particular indicators we employ for detecting teacher cheating, tests the assumptions of the model, and presents simulation results demonstrating the degree to which our approach succeeds in detecting cheating. 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.

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 the pattern of answers given by students in class c are (e.g. is there are unusual blocks of answers, or an especially high degree of correlation across student responses). For simplicity, we assume that these two measures take on one of two values: $SCORE_c = \{\text{low,high}\}$ and $ANSWERS_c = \{\text{typical,unusual}\}$. 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) *Cheating and non-cheating classrooms are drawn from the same underlying distribution of the two outcome measures, SCORE and ANSWERS.*

Thus, if cheating classrooms had not cheated, the probability they would have had a high value on $SCORE$ or an unusual value on $ANSWERS$ is the same as for 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, unconditionally as well as conditional on the other measure:

$$\begin{aligned} & \Pr(SCORE_c = high \mid CHEAT_c = 1, ANSWERS_c) > \\ & \Pr(SCORE_c = high \mid CHEAT_c = 0, ANSWERS_c) \\ (A2) \end{aligned}$$

$$\begin{aligned} & \Pr(ANSWERS_c = unusual \mid CHEAT_c = 1, SCORE_c) > \\ & \Pr(ANSWERS_c = unusual \mid CHEAT_c = 0, SCORE_c) \end{aligned}$$

We define S_{nc} as the probability that a non-cheating class has a high score and A_{nc} as the probability that a non-cheating class has an unusual answer string. 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:

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

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

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

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

Lemma 1 says that the observed 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. If all cheating classrooms have high test scores and unusual answer strings, then the bounds are strict, otherwise they are not.

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 measure of 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:

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

\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 (BUT WE WILL HAVE FALSE POSITIVES TOO??). Second, by Lemma 1, the estimated 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:

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

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 Appendix A

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.⁶ By focusing 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:

⁶We 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.

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.⁷

In practice, the choice of a cutoff for what represents an “unexpectedly” large test score gain or loss is 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,⁸ and construct the following statistic:

$$(3) \quad SCORE_{cbt} = (rank_gain_{c,b,t})^2 + (rank_gain_{c,b,t+1})^2$$

where $rank_gain_{cbt}$ is the percentile rank for class c in subject b in year t . Squaring the individual terms gives more relatively more weight to big test score gains this year and big test score declines the following year.⁹ Classes with relatively big gains on this year’s test and relatively small gains on next year’s test will have high values of $SCORE$. We utilize three different possible cutoffs for this variable in our empirical analysis: classrooms above the 80th, 90th, and 95th percentiles.

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

⁷ Empirically, on average about 60 percent of the excess test score gain in one year is maintained on the next year’s test.

⁸ 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.

⁹ 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.

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 of other students (in the same year, grade and subject) with similar background characteristics choosing that response. Under the assumption that the predicted probability of answering different questions is uncorrelated (in the absence of cheating), we calculate the probability of students answering strings of consecutive questions as they in fact did. Therefore, the probability that a bright child will correctly answer the first five questions on an exam (generally the easier questions) will generally be quite high whereas the probability of any child answering the final five questions on the exam correctly will be quite low. We then search over all combinations of students and consecutive questions to find the block of answers least likely to have arisen by chance.¹⁰ The more unusual is the most unusual block of test responses, the more likely it is that cheating occurred.

¹⁰ Note that we do not require the answers to be correct. Indeed, in many classrooms, the most unusual strings include some incorrect answers.

The second measure of suspicious answer strings involves the overall degree of correlation in student answers across the test. This measure is meant to capture more general patterns of similarity in student responses. For example, when a teacher changes answers on test forms, it presumably increases the uniformity of student test forms across students in the class. 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 variance of student response on each question and take the average over all 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. For example, if the teacher changes answers for multiple students on some 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 taking the identical test and getting 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. If students in a class miss the easy questions while correctly answering the hard questions, this could 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.¹¹

$$(4) \quad ANSWERS_{cbt} = (rank_m1_{c,b,t})^2 + (rank_m2_{c,b,t})^2 + (rank_m3_{c,b,t})^2 + (rank_m4_{c,b,t})^2$$

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

The Cheating Indicators in Practice

Figure 2, which presents student answer strings test scores for two actual classrooms, provides an example of how our cheating indicators work in practice. The top bottom of data in Figure 2 corresponds to a typical classroom; the top panel is a class in which we suspect teacher cheating occurred. 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 the bottom panel of Figure 2, we see that 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. The most unlikely block of answers in the two classrooms (the first measure of suspicious

¹¹ Because different subjects and grades have differing numbers of questions, it is difficult to make meaningful

answers) is identified by a box placed around the responses in Figure 2 [ADD THIS]. The likelihood of that particular string occurring by chance in the top classroom is XX, placing the class in the 99th percentile on this measure.¹² In the second class shown in Figure 2, XX kids give identical answers on XX consecutive questions. This block would be predicted to arise by chance with probability XX, or XX times less frequently than the string in the first classroom.¹³ The bottom classroom ranks in the Xxth 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 places it at the Xxth percentile among classrooms. The bottom class, in contrast, is in the Xxth percentile on this measure. With respect to the variance in within-class correlation across questions, the top class is at the Xxth percentile; the bottom class is at the Xxth.

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

comparisons across tests on the raw indicators.

¹² Because we have searched over every possible combination of students and answers to find the least likely outcome, the number XX has no direct interpretation. It will, however, provide a basis for comparison across classrooms.

¹³ SOME FOOTNOTE TO PUT INTO PERSPECTIVE HOW SMALL THAT NUMBER IS.

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 ranks at the Xxth percentile on this measure, compared to Xxth percentile for the class in the bottom panel. Overall, on our summary measure for suspicious answer strings that combines all four facets, the top classroom ranks at the Xxth percentile and the bottom classroom is at the Xxth.

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 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.¹⁴ 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 1993-1999, which encompasses over 500,000 student-years of data.¹⁵ 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.

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 (XX percent of the sample).¹⁶ Students with missing demographic data (XX percent) are also excluded from the analysis. Finally, because our algorithms for identifying cheating rely on

¹⁴ 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.

¹⁵ 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.

¹⁶ Test data may be missing either because a student did not attend school on the days of the test, or because the student transferred into the CPS system in the current year or left the system prior to the next year of testing.

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 (XX percent). 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 34,000 classroom-years of data (with four subject tests per classroom-year) and over 700,000 student-year observations.

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 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.¹⁷ 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

¹⁷The 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 computation, problem-solving and data analysis.

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 1997.¹⁸

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 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. Nearly one-quarter of elementary school students in Chicago are either exempt from testing or excluded from official test reporting due to placement in bilingual or special education programs.

V. Testing the Assumptions of the Model

Our identification strategy for detecting cheating rests on two assumptions. First, for non-cheating classrooms our two measures of cheating are either uncorrelated (or, alternatively, in a straightforward generalization of the model the degree of correlation is known). Second, cheating classrooms can be identified through the combination of large test score fluctuations and suspicious test strings. In this section, we present evidence on the plausibility of each of these key assumptions.

¹⁸ These three forms are used for re-testing, summer school testing, and mid-year testing as well, so that it is likely

Are the Two Cheating Measures Correlated in Non-Cheating Classrooms?

If we knew with certainty the classrooms that were cheating, then it would be straightforward to test the degree of correlation among non-cheating classrooms.¹⁹ 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, the bottom portion of the distributions will be primarily composed of non-cheating classes. Thus, by focusing on this range, we are able to offer a partial test of the degree of correlation in non-cheating rooms.

The top panel of Table 2 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. These numbers mirror Figure 1. As in the picture, we can see that the likelihood of large test-score fluctuations increases slightly as one moves from left to right, although the gains are relatively minor. The bottom panel of Table 2 reverses the exercise, showing how quartiles of *SCORE* predict a high value of *ANSWERS*. Classes that are in the bottom quartile on *SCORE* have relatively high rates of suspicious strings. For instance, 24.8 percent of such classrooms are above the 80th percentile on *ANSWERS*. The probabilities of suspicious strings for classes in the second and third quartile are similar to one another, and much lower than those of the first quartile.

The key question given Table 2 is what the predicted frequencies would be in the top quartile in the absence of cheating. There are two potential biases at work. First, the values in the table may overstate the baseline rates in non-cheating classes if some cheating classrooms

that over the years, teachers have seen the same exam on a number of occasions.

¹⁹Of course, if we knew which classrooms were cheating, then this assumption would not be necessary in the first place.

fall below the top quartile.²⁰ If cheating classrooms are especially prevalent in the third quartile, than those results will be most elevated. 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 first bias argues against focusing on the third quartile, the second bias argues in favor of using it. Because our greatest concern is falsely labeling honest teachers as cheaters, we use the third quartile as our baseline. Given the similarity of the results across quartiles in Table 2, however, our results are not sensitive to choosing other benchmarks (e.g. the average over the bottom three quartiles, projecting a linear trend using the first three quartiles, etc.).

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

²⁰Cheating classrooms are more likely to be at the top end of the distribution on both measures and are also likely to be positively correlated on the measures, inducing an upward bias.

class.²¹ Also, 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 middle threshold (90th percentile) on both *ANSWERS* and *SCORE*. We present the results for 5th grade reading in 1993. This grade and year were 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 it is not a benchmark grade and the time period is prior to the accountability reforms.

Table 3 reports the results of the simulation. As a point of reference, 1.79 percent of the classrooms in the actual data exceed the thresholds we use for the calculations in the table. For the most minor case of cheating (3 questions for 25 percent of the class), our cheating indicator picks up less than 9 and 6 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 – 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 70

²¹ We have also experimented with forms of cheating with intermediate degrees of sophistication. Not surprisingly,

percent of the unsophisticated cheaters are detected, as well as more than half of the sophisticated cheaters. Note, however, that the likelihood that a good teacher is labeled as a cheater—while still much lower than that of the cheaters (13.65 percent)—also rises. By the most extreme cases we examine (9 questions for 100 percent of the class), virtually every classroom – including the good teachers – is categorized as cheating. 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 less than one XXth of a percent of classrooms in our sample), we will mistakenly label them as cheaters.²²

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 moderate 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.

²² 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.

VI. Estimating the Prevalence of Cheating

Following the model presented earlier, to estimate the prevalence of cheating we compare the number of classrooms that we actually observe in the data exhibiting both large test score fluctuations and suspicious answer strings, relative to what we would expect to observe based on estimated correlations between the two measures among non-cheating classrooms.

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 2.0 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. If one looks at specific subject tests, cheating rates are slightly higher on reading comprehension and the first of the three math tests given; cheating rates are consistently lowest on the second math test (the exam that measures XXX).

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 3.4-5.4 percent of all classrooms.²³

Are We Really Detecting Cheating?

One check on whether what we are detecting is really teacher cheating is to examine the persistence of the 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. Table 5 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 (regardless of how suspicious the classroom's answer strings were).²⁴ These classes are further divided into five groups according to how suspicious their answer string patterns were: less than the 50th percentile, 50th-80th percentile, 80th-95th percentile, 95th-99th, and greater than the 99th percentile. It is important to note that this table differs from our previous analysis in that we are in no way conditioning on the following year's test scores, unlike when we construct our cheating estimates. 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 that is *lost* the following year should increase moving from left to right in the table. For classes whose answer strings are

²³ 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.

²⁴ Results are similar if one looks at the top one percent or five percent of classes.

below the mean in their degree of suspiciousness, only 19 percent of the excess gain disappears the following year—that is, over 80 percent of the gains persist. As one moves from left to right in the table, 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, almost 85 percent of the apparent gains in the current year evaporate in the next year’s test.²⁵

If what we are detecting is not teacher cheating, but rather something unusual about the students in a particular class (e.g. a particular sets of skills, high effort at the beginning of the test but not at the end, cheating by individual students, etc.), one might expect that there would be correlation across years in how suspicious a student’s answer strings look. To test this hypothesis, we rank each student by the most unusual block of answers that particular student is part of (this corresponds to the first of the suspicious string measures we introduced in Section III), relative to all other students in that grade, year, and subject. Then we compute a rank-order correlation of that measure for students over time. RESULTS.²⁶

A further check on our results involves the correlation within classrooms and schools on our cheating measures. For example, if what we are detecting is not cheating, but rather something unusual about a particular exam, or a particular day on which the students took the exam, then one would not expect to find a correlation in our cheating indicators across exams for a specific classroom in a given year, particularly since the math and reading exams are generally given on different days. Insofar as most teachers do not administer the exams to their own

²⁵ 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.

²⁶ Because we do not have teacher identifiers in our data, we cannot do the parallel calculation for teachers. We do, however, have room numbers. The year-to-year correlation in cheating behavior for a particular room number in a given school is high. For instance, the probability that a classroom will be above our medium-medium cheating

children, it is may be more likely that cheating is organized or carried out at a school-wide level, by the test coordinator or other school administrator, or a group of teachers. If this were true, one would expect to find multiple classes in a school cheating in any given year, and perhaps even that cheating in one year predicts cheating in future years.²⁷

To examine these possibilities, Table 6 reports the patterns of correlation within classrooms and schools on our cheating measures. The dependent variable is an indicator that a classroom was above the 90th percentile on both of the cheating measures. The baseline probability of qualifying as a cheater for these two cutoffs is 2.6 percent and 1.3 percent respectively. Column 1 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 – five to ten times higher than if the classroom did not cheat on any of the other subjects (which is the omitted category). Classrooms that cheat on two other subjects have are 30 percentage points more likely to cheat on this test. A class that cheats on all three other subjects is about 57 percentage points more likely to 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 1.5 percentage points, or almost 60 percent. These results suggest

threshold conditional on that same room number being above the threshold in the previous year is XX percent, compared to only XX percent if the classroom was not above the threshold in the previous year.

²⁷ 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.

centralized cheating by a school counselor, test coordinator or the principal, rather than by teachers operating independently.

Cheating in the classroom last year also predicts cheating this year. In column 3, for example, we see that classroom's that cheated in the same subject last year are 8.7 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. Note, however, that we cannot track individual teachers across years – we can only link classrooms. The classroom assignment of a teacher may vary from year to year, and there are also high rates of turnover in the CPS system. Finally, column 4 shows that prior cheating in the school strongly predicts the likelihood that a classroom will cheat this year.

[ADD AUDIT STUFF HERE]

One of the greatest risks in examining the prevalence of cheating is to incorrectly label particularly effective teachers as cheaters, simply because they were able to generate unusually large gains for their students. Our algorithms guard against this by taking into account test score results from the subsequent year and relying on unexpected patterns of test score responses. Another way to explore degree of misclassification is to examine how often a classroom's most suspicious string pattern falls within a single passage on the reading exam or within a single topic area on the math exam. For example, if a teacher spends an entire semester studying the Underground Railroad, and the reading exam that year happens to include a passage by Harriet Tubman, 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 that passage. Similarly, 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 average or worse on the other math questions. ADD RESULTS.

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. School-level results were widely publicized only for third and sixth grade test scores. 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 standardized achievement exams in both reading and

²⁸Math performance was not used in this calculation.

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 90th percentile cutoff—that is, a classroom is designated a cheater if its SCORE and ANSWERS are above the 90th 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 are positively correlated with the likelihood of classroom cheating, although the point estimates are only marginally significant. In addition, cheating appears responsive to other costs and benefits. Classrooms that tested poorly last year are much more likely to cheat. Mixed grade classes are significantly less likely to cheat. 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

²⁹ Seven high schools have been reconstituted to date, although no elementary schools have suffered this fate.

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.

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

³² 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. Also note that the results are not sensitive to the cheating cutoff used.

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 10 percent. Teachers who administer the exam to their own students are 0.75 percentage points—nearly 30 percent—more likely to cheat. Finally, there is no apparent 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).

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 XXX percentage point in response to the school probation policy and roughly XXX percentage points due to the ending of social promotion. The magnitude of these changes are 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. A classroom one standard deviation above the mean does not see any increase 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. Column 3 shows Column 4 includes fixed effects and obtains similar results. FINISH WRITE-UP HERE.

For Whom Do Teachers Cheat?

WRITE UP THE RESULTS FOR TABLE 10

VIII. Conclusions

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. Teacher cheating appears quite responsive to relatively minor changes in incentives.

Our results suggest that implementation of high-stakes testing must be done with some caution. If strong incentives are created without instituting safeguards against cheating, we would expect large to see large increases in teacher (or principal) cheating. The types of safeguards against cheating that we have in mind are relatively inexpensive. For instance, hiring outsiders to proctor the exams rather than the teachers themselves. Or perhaps having teachers from one school administer the tests at another school. Yet, even as such tests have become increasingly high stakes, school systems have failed to adopt such measures.

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 Schargrotsky 2001, Duggan and Levitt 2001). 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.

There typically is no clear outcome variable in the corruption case, making it necessary for the researcher to employ non-standard approaches in generating such a measure.

References

- Cizek, G. J. (1999). *Cheating on Tests: How to Do It, Detect It and Prevent It*. New Jersey: Lawrence Erlbaum Associates.
- Shepard, L.A. and Dougherty, 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).
- 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).
- 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.
- Heubert, J. P. and R. M. Hauser, Eds. (1999). *High Stakes: Testing for Tracking, Promotion and Graduation*. Washington, D.C., National Academy Press.
- Wollack, J.A. (1997). A Nominal Response Model Approach for Detecting Answer Copying. *Applied Psychological Measurement*, 22(2), 144-152.
- Lindsay, D. (1996, October 2). Whodunit? Officials find thousands of erasures on standardized tests and suspect tampering. *Education Week*, 25-29.
- 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.

- Aiiken, 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.
- Frary, R.B. (1993). Statistical detection 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.
- 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.

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.

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

where Y_{isc} indicates the response of student s in class c on item i , the number of possible responses (J) is four, and 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.

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

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:

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

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} as the string for student s , then we can express the product as:

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

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 $3 \leq m - n \leq 7$.

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.

$$(5) \quad \begin{aligned} e_{jisc} &= 0 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^J e^{\hat{\beta}_j x_s}} \text{ if } j \neq k \\ &= 1 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^J e^{\hat{\beta}_j x_s}} \text{ if } j = k \end{aligned}$$

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.

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

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.

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

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

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, \bar{q}_i^A , can be expressed as

$$(8) \quad \bar{q}_i^A = \frac{\sum_{s \in \{z: A_z = A_s\}} q_{isc}}{ns_A}$$

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.

$$(9) \quad Z_{sc} = \sum_i (q_{isc} - \bar{q}_i^A)^2$$

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

$$\mathbf{Measure\ 4:} \quad M4_c = \sum_s (Z_{sc} - \bar{Z}^A)$$

