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

BRIAN A. JACOB AND STEVEN D. LEVITT

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. The observed frequency of cheating appears to respond strongly to relatively minor changes in incentives. Our results highlight the fact that high-powered incentive systems, especially those with bright line rules, may induce unexpected 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.

## I. INTRODUCTION

High-powered incentive schemes are designed to align the behavior of agents with the interests of the principal implementing the system. A shortcoming of such schemes, however, is that they are likely to induce behavior distortions along other dimensions as agents seek to game the rules (see, for instance, Holmstrom and Milgrom [1991] and Baker [1992]). The distortions may be particularly pronounced in systems with bright line rules [Glaeser and Shleifer 2001]. It may be impossible to anticipate the many ways in which a particular incentive scheme may be gamed.

Test-based accountability systems in education provide an excellent example of the costs and benefits of high-powered incentive schemes. In an effort to improve student achievement, a number of states and districts have recently implemented programs that use student test scores to punish or reward schools.

* We would like to thank Suzanne Cooper, Mark Duggan, Susan Dynarski, Arne Duncan, Michael Greenstone, James Heckman, Lars Lefgren, two anonymous referees, the editor, Edward Glaeser, and seminar participants too numerous to mention for helpful comments and discussions. We also thank Arne Duncan, Philip 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. 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.

*The Quarterly Journal of Economics,* August 2003

Recent federal legislation institutionalizes this practice, requiring states to test elementary students each year, rate schools on the basis of student performance, and intervene in schools that do not make sufficient improvement.[1] Several prior studies suggest that such accountability policies may be effective at raising student achievement [Richards and Sheu 1992; Grissmer, Flanagan, et al. 2000; Deere and Strayer 2001; Jacob 2002; Carnoy and Loeb 2002; Hanushek and Raymond 2002]. At the same time, however, researchers have documented instances of manipulation, including documented shifts away from nontested areas or "teaching to the test" [Klein et al. 2002; Jacob 2002], and increasing placement in special education [Jacob 2002; Figlio and Getzler 2002; Cullen and Reback 2002].

In this paper we explore a very different mechanism for inflating test scores: outright 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, 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.[3] While such allegations may seem far-fetched, documented cases of 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].

There has been very little previous empirical analysis of teacher cheating.[4] The few studies that do exist involve investi-

[1] The federal legislation, No Child Left Behind, was passed in 2001. Prior to this legislation, virtually every state had linked test-score outcomes to school funding or required students to pass an exit examination to graduate high school. In the state of California, a policy providing for merit pay bonuses of as much as $25,000 per teacher in schools with large test score gains was recently put into place.

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

[3] 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.

[4] 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; Frary, Tideman, and Watts 1977; Angoff 1974]. These methods involve the identification of unusual

gations of specific instances of cheating and generally rely on the analysis of erasure patterns and the controlled retesting of students.[5] 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) that includes the question-by-question answers given by every student in grades 3 to 8 who took the Iowa Test of Basic Skills (ITBS) from 1993 to 2000.[6] 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, and extensive demographic and socioeconomic characteristics.

Our approach to detecting classroom cheating combines 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

patterns of agreement in student responses and, for the most part, are only effective in identifying the most egregious cases of copying.

[5] 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 was 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 retesting resulted in significantly lower scores [Lindsay 1996].

[6] 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.

questions correctly while missing many simple questions). Our identification strategy exploits the fact that these two types of indicators are very weakly correlated in classrooms unlikely to have cheated, but very highly correlated in situations where cheating likely occurred. That allows us to credibly estimate the prevalence of cheating without having to invoke arbitrary cutoffs as to what constitutes cheating.

Empirically, we detect cheating in approximately 4 to 5 percent of the classes in our sample. This estimate is likely to understate the true incidence of cheating for two reasons. First, we focus only on the most egregious type of cheating, where teachers systematically alter 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. As discussed later, 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 by our methods.

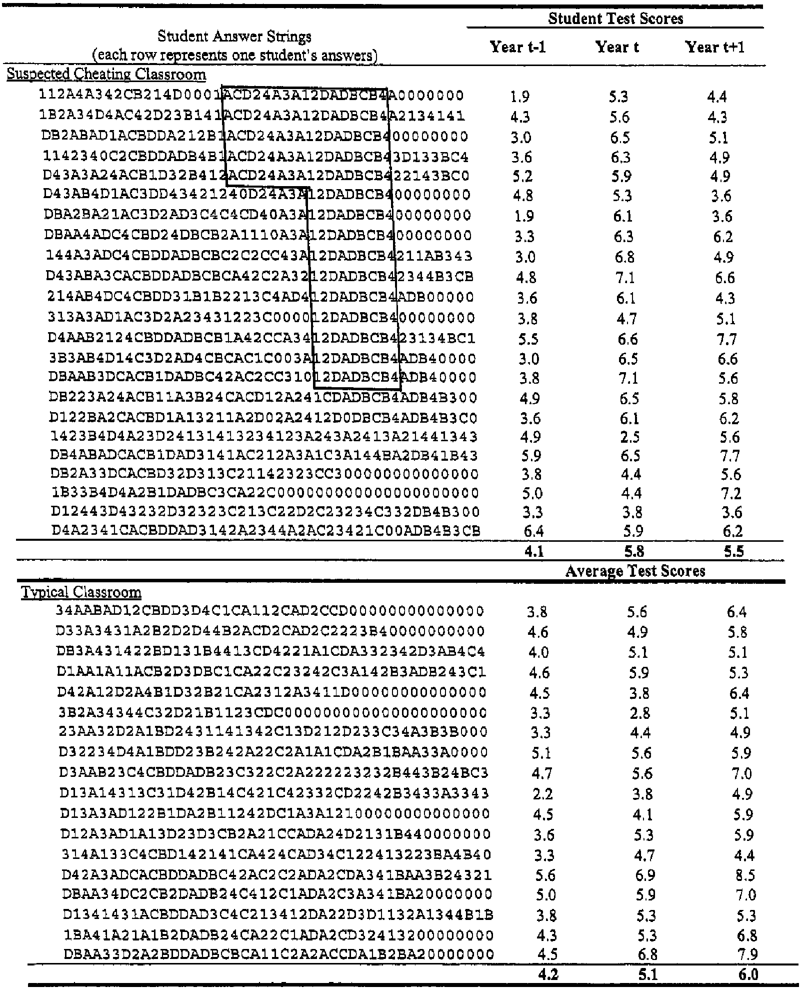
A number of patterns in the results reinforce our confidence that what we measure is indeed cheating. First, simulation results demonstrate that there is nothing mechanical about our identification approach that automatically generates patterns like those observed in the data. When we randomly assign students to classrooms and search for cheating in these simulated classes, our methods find little evidence of cheating. Second, 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). Third, cheating is also correlated within classrooms over time and across classrooms in a particular school. Finally, and perhaps most convincingly, with the cooperation of the Chicago public schools we were able to conduct a *prospective* test of our methods in which we retested a subset of classrooms under controlled conditions that precluded teacher cheating. Classrooms identified as likely cheaters experienced large declines in test scores on the retest, whereas classrooms not suspected of cheating maintained their test score gains.

The prevalence of cheating is also shown 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, particularly for low-achieving students. Following the introduction of these policies, the prevalence of cheating rose sharply in low-achieving classrooms, whereas classes with average or higher-achieving students showed no increase in cheating. Cheating prevalence also 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 discusses the set of indicators we use to capture cheating behavior. Section III describes our identification strategy. Section IV provides a brief overview of the institutional details of the Chicago public schools and the data set that we use. Section V reports the basic empirical results on the prevalence of cheating and presents a wide range of evidence supporting the interpretation of these results as cheating. Section VI analyzes how teacher cheating responds to incentives. Section VII discusses the results and the implications for increasing reliance on high-stakes testing.

## II. Indicators of Teacher Cheating

Teacher cheating, especially in extreme cases, is likely to leave tell-tale signs. In motivating our discussion of the indicators we employ for detecting cheating, it is instructive to compare two actual classrooms taking the same test (see Figure I). Each row in Figure I 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. Test scores are in units of "grade equivalents." The grading scale is normed so that a student at the national average for sixth graders taking the test in the eighth month of the school year would score 6.8. A typical



| Student Answer Strings<br>(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 |
| 114234OC2CBDDADB4B1ACD24A3A12DADBCB43D133BC4 | 3.6 | 6.3 | 4.9 |
| D43A3A24ACB1D32B412ACD24A3A12DADBCB422143BC0 | 5.2 | 5.9 | 4.9 |
| D43AB4D1AC3DD43421240D24A3A12DADBCB400000000 | 4.8 | 5.3 | 3.6 |
| 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 |
| 1B33B4D4A2B1DADBC3CA22C000000000000000000000 | 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** |
| | Average Test Scores | | |
| Typical Classroom | | | |
| 34AABAD12CBDD3D4C1CA112CAD2CCD00000000000000 | 3.8 | 5.6 | 6.4 |
| D33A3431A2B2D2D44B2ACD2CAD2C2223B40000000000 | 4.6 | 4.9 | 5.8 |
| DB3A431422BD131B4413CD4221A1CDA332342D3AB4C4 | 4.0 | 5.1 | 5.1 |
| D1AA1A11ACB2D3DBC1CA22C23242C3A142B3ADB243C1 | 4.6 | 5.9 | 5.3 |
| D42A12D2A4B1D32B21CA2312A3411D00000000000000 | 4.5 | 3.8 | 6.4 |
| 3B2A34344C32D21B1123CDC000000000000000000000 | 3.3 | 2.8 | 5.1 |
| 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** |

FIGURE I

Sample Answer Strings and Test Scores from Two Classrooms

The data in the figure 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 figure. 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.

student would be expected to gain one grade equivalent for each year of school.

The top panel of data shows a class in which we suspect teacher cheating took place; the bottom panel corresponds to a typical classroom. Two striking differences between the classrooms are readily apparent. First, in the cheating classroom, a large block of students provided identical answers on consecutive questions in the middle of the test, as indicated by the boxed area in the figure. For the other classroom, no such pattern exists. Second, looking at the pattern of test scores, students in the cheating classroom experienced large increases in test scores from the previous to the current year (1.7 grade equivalents on average), and actually experienced *declines* on average the following year. In contrast, the students in the typical classroom gained roughly one grade equivalent each year, as would be expected.

The indicators we use as evidence of cheating formalize and extend the basic picture that emerges from Figure I. We divide these indicators into two distinct groups that, respectively, capture unusual test score fluctuations and suspicious patterns of answer strings. In this section we describe informally the measures that we use. A more rigorous treatment of their construction is provided in the Appendix.

### *III.A. Indicator One: Unexpected Test Score Fluctuations*

Given that the aim of cheating is to raise test scores, one signal of teacher cheating is an unusually large gain 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 unusually small 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. We construct a summary measure of how unusual the test score fluctuations are by ranking each classroom's average test score gains relative to all other classrooms in that same subject, grade, and year,[7] and computing 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 relatively more weight to big test score gains this year and big test score declines the following year.[8] In the empirical analysis we consider three possible cutoffs for what it means to have a "high" value on *SCORE*, corresponding to the eightieth, ninetieth, and ninety-fifth percentiles among all classrooms in the sample.

### III.B. Indicator Two: Suspicious 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, as was apparently done in the classroom in the top panel of Figure I. 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. 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

[7] 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.

[8] 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.

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.[9] 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 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 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. If the teacher changes answers for multiple students on selected questions, the within-class correlation on

[9] 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.

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 with the answers of other students in the system who take the identical test and get the exact same score. This measure relies on the fact that questions vary significantly in difficulty. The typical student will answer most of the easy questions correctly, but get many 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:[10]

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

In the empirical work, we again use three possible cutoffs for potential cheating: eightieth, ninetieth, and ninety-fifth percentiles.

## III. A Strategy for Identifying the Prevalence of Cheating

The previous section described indicators that are likely to be correlated with cheating. Because sometimes such patterns arise by chance, however, not every classroom with large test score fluctuations and suspicious answer strings is cheating. Furthermore, the particular choice of what qualifies as a "large" fluctuation or a "suspicious" set of answer strings will necessarily be arbitrary. In this section we present an identification strategy that, under a set of defensible assumptions, nonetheless provides estimates of the prevalence of cheating.

To identify the number of cheating classrooms in a given

[10] Because different subjects and grades have differing numbers of questions, it is difficult to make meaningful comparisons across tests on the raw indicators.

year, we would like to compare the observed joint distribution of test score fluctuations and suspicious answer strings with a counterfactual distribution in which no cheating occurs. Differences between the two distributions would provide an estimate of how much cheating occurred. If teachers in a particular school or year are cheating, there will be more classrooms exhibiting both unusual test score fluctuations and suspicious answer strings than otherwise expected. In practice, we do not have the luxury of observing such a counterfactual. Instead, we must 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 both large test score fluctuations and suspicious answer strings, (2) if cheating classrooms had not cheated, their distribution of test score fluctuations and answer strings patterns would be identical to noncheating classrooms, and (3) in noncheating classrooms, the correlation between test score fluctuations and suspicious answers is constant throughout the distribution.[11]

If assumption (1) holds, then cheating classrooms will be concentrated in the upper tail of the joint distribution of unusual test score fluctuations and suspicious answer strings. Other parts of the distribution (e.g., classrooms ranked in the fiftieth to seventy-fifth percentile of suspicious answer strings) will consequently include few cheaters. As long as cheating classrooms would look similar to noncheating classrooms on our measures if they had not cheated (assumption (2)), classes in the part of the distribution with few cheaters provide the noncheating counterfactual that we are seeking.

The difficulty, however, is that we only observe this noncheating counterfactual in the bottom and middle parts of the distribution, when what we really care about is the upper tail of the distribution where the cheaters are concentrated. Assumption (3), which requires that in noncheating classrooms the correlation between test score fluctuations and suspicious answers is constant throughout the distribution, provides a potential solution. If this assumption holds, we can use the part of the distribution that is relatively free of cheaters to project what the right tail of the distribution would look like absent cheating. The gap between the predicted and observed frequency of classrooms that

[11] We formally derive the mathematical model described in this section in Jacob and Levitt [2003].

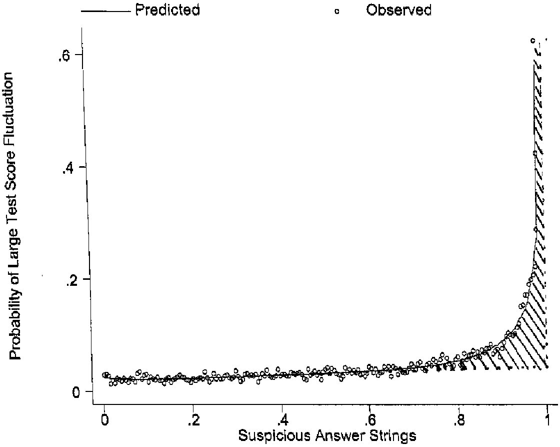


FIGURE II

The Relationship between Unusual Test Scores and Suspicious Answer Strings

The horizontal axis reflects a classroom's percentile rank in the distribution of suspicious answer strings within a given grade, subject, and year, with zero representing the least suspicious classroom and one representing the most suspicious classroom. The vertical axis is the probability that a classroom will be above the ninety-fifth percentile on our measure of unusual test score fluctuations. The circles in the figure 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.

are extreme on both the test score fluctuation and suspicious answer string measures provides our estimate of cheating.

Figure II demonstrates how our identification strategy works empirically. The horizontal axis in the figure ranks classrooms according to how suspicious their answer strings are. The vertical axis is the fraction of the classrooms that are above the ninety-fifth percentile on the unusual test score fluctuations measure.[12] The graph combines all classrooms and all subjects in our data.[13] Over most of the range (roughly from zero to the seventy-fifth percentile on the horizontal axis), there is a slight positive correlation between unusual test scores and suspicious answer strings.

[12] The choice of the ninety-fifth percentile is somewhat arbitrary. In the empirical work that follows we also consider the eightieth and ninetieth percentiles. The choice of the cutoff does not affect the basic patterns observed in the data.

[13] 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.

This is the part of the distribution that is unlikely to contain many cheating classrooms. Under the assumption that the correlation between our two measures is constant in noncheating classrooms over the whole distribution, we would predict that the straight line observed for all but the right-hand portion of the graph would continue all the way to the right edge of the graph.

Instead, as one approaches the extreme right tail of the distribution of suspicious answer strings, the probability of large test score fluctuations rises dramatically. That sharp rise, we argue, is a consequence of cheating. To estimate the prevalence of cheating, we simply compare the actual area under the curve in the far right tail of Figure I to the predicted area under the curve in the right tail projecting the linear relationship observed from the fiftieth to seventy-fifth percentile out to the ninety-ninth percentile. Because this identification strategy is necessarily indirect, we devote a great deal of attention to presenting a wide variety of tests of the validity of our approach, the sensitivity of the results to alternative assumptions, and the plausibility of our findings.

## 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 1993–2000.[15] For each student we have the question-by-question answer string on each year's 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 a wide

[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 also have test scores for eighth graders, but we exclude them because our algorithm requires test score data for the following year and the ITBS test is not administered to ninth graders.

range of school-level characteristics. 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 (in addition to those with missing test scores in the baseline year).[16] Less than one-half of 1 percent of 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 (roughly 3 percent of students). A handful of classrooms recorded as having more than 40 students—presumably multiple classrooms not separately identified in the data—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 Table I where the unit of observation is at the level of class*subject*year. As in many urban school districts, students in Chicago are disproportionately poor and minority. Roughly 60 percent of students are Black, and 26 percent are Hispanic, with nearly 85 percent receiving free or reduced-price lunch. Less than 29 percent of students scored at or above national norms in reading.

The ITBS exams are administered over a weeklong 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

[16] The exact number of students with missing test data varies by year and grade. Overall, we drop roughly 12 percent of students because of missing test score information in the baseline year. These are students who (i) were not tested because of placement in bilingual or special education programs or (ii) simply missed the exam that particular year. In addition to these students, we also drop approximately 13 percent of students who were missing test scores in either the preceding or following years. 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.

TABLE I
SUMMARY STATISTICS

| | Mean<br>(sd) |
|---|---|
| **Classroom characteristics** | |
| Mixed grade classroom | 0.073 |
| Teacher administers exams to her own students (3rd grade) | 0.206 |
| Percent of students who were tested and included in official reporting | 0.883 |
| Average prior achievement | −0.004 |
| (as deviation from year*grade*subject mean) | (0.661) |
| % Black | 0.595 |
| % Hispanic | 0.263 |
| % Male | 0.495 |
| % Old for grade | 0.086 |
| % Living in foster care | 0.044 |
| % Living with nonparental relative | 0.104 |
| Cheater—95th percentile cutoff | 0.013 |
| **School characteristics** | |
| Average quality of teachers' undergraduate institution | −2.550 |
| in the school | (0.877) |
| Percent of teachers who live in Chicago | 0.712 |
| Percent of teachers who have an MA or a PhD | 0.475 |
| Percent of teachers who majored in education | 0.712 |
| Percent of teachers under 30 years of age | 0.114 |
| Percent of teachers at the school less than 3 years | 0.547 |
| % students at national norms in reading last year | 28.8 |
| % students receiving free lunch in school | 84.7 |
| Predominantly Black school | 0.522 |
| Predominantly Hispanic school | 0.205 |
| Mobility rate in school | 28.6 |
| Attendance rate in school | 92.6 |
| School size | 722 |
| | (317) |
| **Accountability policy** | |
| Social promotion policy | 0.215 |
| School probation policy | 0.127 |
| Test form offered for the first time | 0.371 |
| Number of observations | 163,474 |

The data summarized above include one observation per classroom*year*subject. The sample includes all students in third to seventh grade for the years 1993–2000. We drop observations for the following reasons: (a) missing reading or math scores in either the baseline year, the preceding year, or the following year; (b) missing demographic data; (c) classrooms for which we have fewer than ten valid students in a particular grade or more than 40 students. See the discussion in the text for a more detailed explanation of the sample, including the percentages excluded for various reasons.

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 consists 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. There is no penalty for guessing. 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 grade equivalent.

After collecting student exams, 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 testing week, the test coordinators at each school deliver the completed answer keys and exams to the CPS central office. School personnel are not supposed 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]

## V. The Prevalence of Teacher Cheating

As noted earlier, our estimates of the prevalence of cheating are derived from a comparison of the actual number of classrooms that are above some threshold on both of our cheating indicators, relative to the number we would expect to observe based on the correlation between these two measures in the 50th–75th percentiles of the suspicious string measure. The top panel of Table II 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.[19] Because

[17] 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 math exam consists of three sections that assess number concepts, problem-solving, and computation.

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

[19] It is important to note that we exclude a particular form of cheating that appears to be quite prevalent in the data: teachers randomly filling in answers left blank by students at the end of the exam. In many classrooms, almost every student will end the test with a long string of the identical answers (typically all the same response like "B" or "D.") 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

TABLE II
ESTIMATED PREVALENCE OF TEACHER CHEATING

| Cutoff for suspicious answer strings (ANSWERS) | Cutoff for test score fluctuations (SCORE): | | |
|---|---|---|---|
| | 80th percentile | 90th percentile | 95th percentile |
| | *Percent cheating on a particular test* | | |
| 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* | | |
| 80th percentile | 4.5 | 5.6 | 5.3 |
| 90th percentile | 4.2 | 4.9 | 4.4 |
| 95th percentile | 3.5 | 3.8 | 3.4 |

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 counted only once. Our sample includes over 35,000 3rd–7th grade classrooms in the Chicago public schools for the years 1993–1999.

the decision as to what cutoff signifies a "high" value on our cheating indicators is arbitrary, we present a 3 × 3 matrix of estimates using three different thresholds (eightieth, ninetieth, and ninety-fifth percentiles) for each of our cheating measures. 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 set of cutoffs, the range of estimates is relatively tight.

The bottom panel of Table II 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, then the results in the bottom panel

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.

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.[20] Because of the necessarily indirect nature of our identification strategy, we explore a range of supplementary tests designed to assess the validity of the estimates.[21]

### V.A. *Simulation Results*

Our prevalence estimates may be biased downward or upward, depending on the extent to which our algorithm fails to successfully identify all instances of cheating ("false negatives") versus the frequency with which we wrongly classify a teacher as cheating ("false positive").

One simple simulation exercise we undertook with respect to possible false positives involves randomly assigning students to hypothetical classrooms. These synthetic classrooms thus consist of students who in actuality had no connection to one another. We then analyze 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 negative 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, there does not appear to be anything mechanical in our algorithm that generates evidence of cheating.

A second simulation involves artificially altering a class-

[20] Computation of the overall prevalence is somewhat 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. Details on these calculations are available from the authors.

[21] In an earlier version of this paper [Jacob and Levitt 2003], we present a number of additional sensitivity analyses that confirm the basic results. First, we test whether our results are sensitive to the way in which we predict the prevalence of high values of ANSWERS and SCORE in the upper part of the other distribution (e.g., fitting a linear or quadratic model to the data in the lower portion of the distribution versus simply using the average value from the 50th–75th percentiles). We find our results are extremely robust. Second, we examine whether our results might simply be due to mean reversion by testing whether classrooms with suspicious answer strings are less likely to maintain large test score gains. Among a sample of classrooms with large test score gains, we find that mean reversion is substantially greater among the set of classes with highly suspicious answer strings. Third, we investigate whether we might be detecting student rather than teacher cheating by examining whether students who have suspicious answer strings in one year are more likely to have suspicious answer strings in other years. We find that they do not.

room's answer strings in ways that we believe mimic teacher cheating or outstanding teaching.[22] We are then able to see how frequently we label these altered classes as "cheating" and how this varies with the extent and nature of the changes we make to a classroom's answers. 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. The outcome of this simulation reveals that our methods are surprisingly weak in detecting moderate cheating, even in the naïve case. For instance, when three consecutive answers are changed to correct for 25 percent of a class, we catch such behavior only 4 percent of the time. Even when six questions are changed for half of the class, we catch the cheating in less than 60 percent of the cases. Only when the cheating is very extreme (e.g., six answers changed for the entire class) are we very likely to catch the cheating. The explanation for this poor performance is that classrooms with low test-score gains, even with the boost provided by this cheating, do not have large enough test score fluctuations to make them appear unusual because there is so much inherent volatility in test scores. When the cheating takes the more sophisticated form described above, we are even less successful at catching low and moderate cheaters (2 percent and 34 percent, respectively, in the first two scenarios described above), but we almost always detect very extreme cheating.

From a political or equity perspective, however, we may be even more concerned with false positives, specifically the risk of accusing highly effective teachers of cheating. Hence, we also simulate the impact of a good teacher, changing certain answers for certain students from incorrect to correct responses. Our manipulation in this case differs from the cheating simulations above in two ways: (1) we allow some of the gains to be preserved in the following year; and (2) we alter questions such that 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. These simulation results suggest

[22] See Jacob and Levitt [2003] for the precise details of this simulation exercise.

that only rarely are good teachers mistakenly labeled as cheaters. For instance, a teacher whose prowess allows him or her to raise test scores by six questions for half the students in the class (more than a 0.5 grade equivalent increase on average for the class) is labeled a cheater only 2 percent of the time. In contrast, a naïve cheater with the same test score gains is correctly identified as a cheater in more than 50 percent of cases.

In another test for false positives, we explored whether we might be mistaking emphasis on certain subject material for cheating. For example, 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. One might imagine a similar scenario in reading if, for example, a teacher creates a lengthy unit on the Underground Railroad, which later happens to appear in a passage on the reading comprehension exam. To examine this, we analyze the nature of the across-question correlations in cheating versus noncheating classrooms. We find that the most highly correlated questions in cheating classrooms are no more likely to measure the same skill (e.g., fractions, geometry) than in noncheating classrooms. The implication of these results is that the prevalence estimates presented above are likely to substantially understate the true extent of cheating—there is little evidence of false positives—but we frequently miss moderate cases of cheating.

### *V.B. The Correlation of Cheating Indicators across Subjects, Classrooms, and Years*

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.

TABLE III
PATTERNS OF CHEATING WITHIN CLASSROOMS AND SCHOOLS

| | Dependent variable = Class suspected of cheating (mean of the dependent variable = 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 | 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) |
| Fixed effects for grade*subject*year | Yes | Yes | Yes | Yes |
| $R^2$ | 0.090 | 0.093 | 0.109 | 0.109 |
| Number of observations | 165,578 | 165,578 | 94,182 | 94,170 |

The dependent variable is an indicator for whether a classroom is above the 95th percentile on both our suspicious strings and unusual test score measures of cheating on a particular subject test. Estimation is done using a linear probability model. The unit of observation is classroom*grade*year*subject. For columns that include measures of cheating in prior years, observations where that classroom 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 III 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 ninety-fifth 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.

Column 1 of Table III 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 affect 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. Even given this important limitation, cheating in the classroom last year predicts cheating this year. In column 3, for example, we see that classrooms 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.

### *V.C. Evidence from Retesting under Controlled Circumstances*

Perhaps the most compelling evidence for our methods comes from the results of a unique policy intervention. In spring 2002 the Chicago public schools provided us the opportunity to retest over 100 classrooms under controlled circumstances that precluded cheating, a few weeks after the initial ITBS exam was administered.[23] Based on suspicious answer strings and unusual test score gains on the initial spring 2002 exam, we identified a set of classrooms most likely to have cheated. In addition, we chose a second group of classrooms that had suspicious answer strings but did not have large test score gains, a pattern that is consistent with a bad teacher who attempts to mask a poor teaching performance by cheating. We also retested two sets of classrooms as control groups. The first control group consisted of classes with large test score gains but no evidence of cheating in their answer strings, a potential indication of effective teaching. These classes would be expected to maintain their gains when retested, subject to possible mean reversion. The second control group consisted of randomly selected classrooms.

The results of the retests are reported in Table IV. Columns 1–3 correspond to reading; columns 4–6 are for math. For each subject we report the test score gain (in standard score units) for three periods: (i) from spring 2001 to the initial spring 2002 test; (ii) from the initial spring 2002 test to the retest a few weeks later; and (iii) from spring 2001 to the spring 2002 retest. For purposes of comparison we also report the average gain for all classrooms in the system for the first of those measures (the other two measures are unavailable unless a classroom was retested).

Classrooms prospectively identified as "most likely to have cheated" experienced gains on the initial spring 2002 test that were nearly twice as large as the typical CPS classroom (see columns 1 and 4). On the retest, however, those excess gains completely disappeared—the gains between spring 2001 and the spring 2002 retest were close to the systemwide average. Those classrooms prospectively identified as "bad teachers likely to have cheated" also experienced large score declines on the retest ($-8.8$ and $-10.5$ standard score points, or roughly two-thirds of a grade equivalent). Indeed, comparing the spring 2001 results with the

[23] Full details of this policy experiment are reported in Jacob and Levitt [2003]. The results of the retests were used by CPS to initiate disciplinary action against a number of teachers, principals, and staff.

TABLE IV
RESULTS OF RETESTS: COMPARISON OF RESULTS FOR SPRING 2002 ITBS AND AUDIT TEST

| | Reading gains between . . . | | | Math gains between . . . | | |
|---|---|---|---|---|---|---|
| Category of classroom | Spring 2001 and spring 2002 | Spring 2001 and 2002 retest | Spring 2002 and 2002 retest | Spring 2001 and spring 2002 | Spring 2001 and 2002 retest | Spring 2002 and 2002 retest |
| All classrooms in the Chicago public schools | 14.3 | — | — | 16.9 | — | — |
| Classrooms prospectively identified as most likely cheaters (N = 36 on math, N = 39 on reading) | 28.8 | 12.6 | −16.2 | 30.0 | 19.3 | −10.7 |
| Classrooms prospectively identified as bad teachers suspected of cheating (N = 16 on math, N = 20 on reading) | 16.6 | 7.8 | −8.8 | 17.3 | 6.8 | −10.5 |
| Classrooms prospectively identified as good teachers who did not cheat (N = 17 on math, N = 17 on reading) | 20.6 | 21.1 | +0.5 | 28.8 | 25.5 | −3.3 |
| Randomly selected classrooms (N = 24 overall, but only one test per classroom) | 14.5 | 12.2 | −2.3 | 14.5 | 12.2 | −2.3 |

Results in this table compare test scores at three points in time (spring 2001, spring 2002, and a retest administered under controlled conditions a few weeks after the initial spring 2002 test). The unit of observation is a classroom-subject test pair. Only students taking all three tests are included in the calculations. Because of limited data, math and reading results for the randomly selected classrooms are combined. Only the first two columns are available for all CPS classrooms since audits were performed only on a subset of classrooms. All entries in the table are in standard score units.

spring 2002 retest, children in these classrooms had gains only half as large as the typical yearly CPS gain. In stark contrast, classrooms prospectively identified as having "good teachers" (i.e., classrooms with large gains but not suspicious answer strings) actually scored even *higher* on the reading retest than they did on the initial test. The math scores fell slightly on retesting, but these classrooms continued to experience extremely large gains. The randomly selected classrooms also maintained almost all of their gains when retested, as would be expected. In summary, our methods proved to be extremely effective both in identifying likely cheaters and in correctly classifying teachers who achieved large test score gains in a legitimate manner.

## VI. 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 responds 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 involved putting schools "on probation" if less than 15 percent of students scored at or above national norms on the ITBS reading exams. The CPS did not use math performance to determine probation status. 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.[24] 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

[24] For a more detailed analysis of the probation policy, see Jacob [2002] and Jacob and Lefgren [forthcoming].

mathematics in order to be promoted to the next grade. The promotion standards were implemented in spring 1997 for third and sixth graders. Promotion decisions are based solely on scores in reading comprehension and mathematics.

Table V presents OLS estimates of the relationship between teacher cheating and a variety of classroom and school characteristics.[25] The unit of observation is a classroom*subject*grade* year. The dependent variable is an indicator of whether we suspect the classroom cheated using our most stringent definition: a classroom is designated a cheater if both its test score fluctuations and the suspiciousness of its answer strings are above the ninety-fifth percentile for that grade, subject, and year.[26]

In column (1) the policy changes are restricted to have a constant impact across all classrooms. 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. 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 responsive to the change in incentives. In column (2) we see that a classroom one standard deviation below the mean increased its likelihood of 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 a baseline rate of cheating of 1.1 percent, these effects are substantial. The magnitude of these changes is particularly large considering that no elementary school on probation was actually reconstituted during this period, and that the social promotion policy has a direct impact on students, but not direct ramifications for teacher pay or rewards. 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. These findings are robust to

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

[26] 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, the primary impact will likely be to simply decrease the precision of our estimates.

TABLE V
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[a] | −0.0007 | −0.0007 | −0.0011 | — |
| | (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 of 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 |

The unit of observation is classroom*grade*year*subject, and the sample includes 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 ninety-fifth percentile cutoff. Robust standard errors clustered by school*year are shown in parentheses. 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 nonparental 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 master's 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.

adding a number of classroom and school characteristics (column (3)) as well as school*year fixed effects (column (4)).

Cheating also appears 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 whose test scores "count" will increase the likelihood of cheating by roughly 20 percent.[27] Teachers who administer the exam to their own students are 0.67 percentage points—approximately 50 percent—more likely to cheat.[28]

A few other notable results emerge. First, 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). Classrooms in schools with lower achievement, higher poverty rates, and more Black students are more likely to cheat. Classrooms in schools with teachers who graduated from more prestigious undergraduate institutions are less likely to cheat, whereas classrooms in schools with younger teachers are more likely to cheat.

[27] Consistent with this finding, within cheating classrooms, teachers are less likely to alter the answer sheets for students who are excluded from test reporting. Teachers also appear to less frequently cheat for boys and for older students.

[28] In an earlier version of the paper [Jacob and Levitt 2002], we examine for which students teachers tend to cheat. We find that teachers are most likely to cheat for students in the second quartile of the national ability distribution (twenty-fifth to fiftieth percentile) in the previous year, consistent with the incentives under the probation policy.

## VII. 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 thousands of instances of classroom cheating, representing 4–5 percent of the classrooms each year. Moreover, we find that teacher cheating appears quite responsive to relatively minor changes in incentives.

The obvious benefits of high-stakes tests as a means of providing incentives must therefore be weighed against the possible distortions that these measures induce. Explicit cheating is merely one form of distortion. Indeed, the kind of cheating we focus on is one that could be greatly reduced at a relatively low cost through the implementation of proper safeguards, such as is done by Educational Testing Service on the SAT or GRE exams.[29]

Even if this particular form of cheating were eliminated, however, there are a virtually infinite number of dimensions along which strategic school personnel can act to game the current system. For instance, Figlio and Winicki [2001] and Rohlfs [2002] document that schools attempt to increase the caloric intake of students on test days, and disruptive students are especially likely to be suspended from school during official testing periods [Figlio 2002]. It may be prohibitively costly or even impossible to completely safeguard the system.

Finally, this paper fits into a small but growing body of research focused on identifying corrupt or illicit behavior on the part of economic actors (see Porter and Zona [1993], Fisman [2001], Di Tella and Schargrodsky [2001], and Duggan and Levitt [2002]). Because individuals engaged in such behavior actively attempt to hide 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

[29] Although even tests administered by the Educational Testing Service are not immune to cheating, as evidenced by recent reports that many of the GRE scores from China are fraudulent [Contreras 2002].

researcher to employ nonstandard approaches in generating such a measure.

## Appendix: The Construction of Suspicious String Measures

This appendix describes in greater detail how we construct each of our measures of unexpected or suspicious responses.

The first measure focuses on the most unlikely block of identical answers given on consecutive questions. 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:

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

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

$$(2)\quad p_{isc} = \frac{e^{\hat{\beta}_k x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{ for } k$$

$$= \text{response actually chosen by student } s \text{ on item } i.$$

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:

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

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 $\bar{S}_{sc}^{mn}$ and as the string for student $s$, then we can express the product as

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

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 $3 \leq m - n \leq 7$. To create our first indicator of suspicious string patterns, we take the minimum of the predicted block probability for each classroom.

MEASURE 1. $\mathrm{M1}_c = \min_s(\tilde{p}_{sc}^{mn})$.

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

$$
\text{(5)} \qquad e_{jisc} = 0 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{ if } j \neq k; \\
= 1 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{ if } j = k,
$$

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:

$$
\text{(6)} \qquad e_{jic} = \sum_{s} e_{jisc}.
$$

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:

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

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. $\mathrm{M2}_c = \bar{v}_c = \sum_i v_{ic}/ni$, where $ni$ is the number of items on the exam.

Our third measure is simply the *variance* (as opposed to the mean) in the degree of correlation across questions within a classroom:

MEASURE 3. $\mathrm{M3}_c = \sigma^2_{v_c} = \Sigma_i\ (v_{ic} - \bar{v}_c)^2/ni.$

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. 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 number of students with an aggregate score of $A$, then this fraction, $\bar{q}^A_i$, can be expressed as

$$\bar{q}^A_i = \frac{\sum_{s\in\{z:A_z=A_s\}} q_{isc}}{ns_A}. \tag{8}$$

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:

$$Z_{sc} = \sum_i (q_{isc} - \bar{q}^A_i)^2. \tag{9}$$

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:

MEASURE 4. $\mathrm{M4}_c = \Sigma_s\ (Z_{sc} - \bar{Z}^A).$

KENNEDY SCHOOL OF GOVERNMENT, HARVARD UNIVERSITY
UNIVERSITY OF CHICAGO AND AMERICAN BAR FOUNDATION

## REFERENCES

Aiken, Linda R., "Detecting, Understanding and Controlling for Cheating on Tests," *Research in Higher Education,* XXXII (1991), 725–736.

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

Baker, George, "Incentive Contracts and Performance Measurement," *Journal of Political Economy,* C (1992), 598–614.

Bellezza, Francis S., and Suzanne F. Bellezza, "Detection of Cheating on Multiple-Choice Tests by Using Error Similarity Analysis," *Teaching of Psychology,* XVI (1989), 151–155.

Carnoy, Martin, and Susanna Loeb, "Does External Accountability Affect Student

Outcomes? A Cross-State Analysis," Working Paper, Stanford University School of Education, 2002.

Contreras, Russell, "Inquiry Uncovers Possible Cheating on GRE in Asia," Associated Press August 7, 2002.

Cullen, Julie Berry, and Randall Reback, "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." Working paper, University of Michigan, 2002.

Deere, Donald, and Wayne Strayer, "Putting Schools to the Test: School Accountability, Incentives and Behavior," Working Paper, Texas A&M University, 2002.

DiTella, Rafael, and Ernesto Schargrodsky, "The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires," unpublished manuscript, Harvard Business School, 2001.

Duggan, Mark, and Steven Levitt, "Winning Isn't Everything: Corruption in Sumo Wrestling," *American Economic Review,* XCII (2002), 1594–1605.

Figlio, David N., "Testing, Crime and Punishment," Working Paper University of Florida, 2002.

Figlio, David N., and Joshua Winicki, "Food for Thought: The Effects of School Accountability Plans on School Nutrition," Working Paper, University of Florida, 2001.

Figlio, David N., and Lawrence S. Getzler, "Accountability, Ability and Disability: Gaming the System?" Working Paper University of Florida, 2002.

Fisman, Ray, "Estimating the Value of Political Connections,"*American Economic Review,* XCI (2001), 1095–1102.

Frary, Robert B., "Statistical Detection of Multiple-Choice Answer Copying: Review and Commentary," *Applied Measurement in Education,* VI (1993), 153–165.

Frary, Robert B., T. Nicholas Tideman, and Thomas Watts, "Indices of Cheating on Multiple-Choice Tests," *Journal of Educational Statistics,* II (1977), 235–256.

Glaeser, Edward, and Andrei Shleifer, "A Reason for Quantity Regulation,"*American Economic Review,* XCI (2001), 431–435.

Grissmer, David W., Ann Flanagan, et al., "Improving Student Achievement: What NAEP Test Scores Tell Us," RAND Corporation, MR-924-EDU, 2000.

Hanushek, Eric A., and Margaret E. Raymond, "Improving Educational Quality: How Best to Evaluate Our Schools?" Conference paper, Federal Reserve Bank of Boston, 2002.

Hofkins, Diane, "Cheating 'Rife' in National Tests," *New York Times Educational Supplement,* June 16, 1995.

Holland, Paul W., "Assessing Unusual Agreement between the Incorrect Answers of Two Examinees Using the K-Index: Statistical Theory and Empirical Support," Technical Report No. 96-5, Educational Testing Service, 1996.

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

Jacob, Brian A., "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," Working Paper No. 8968, National Bureau of Economic Research, 2002.

Jacob, Brian A., and Lars Lefgren, "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago," *Journal of Human Resources,* forthcoming.

Jacob, Brian A., and Steven Levitt, "Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory," Working Paper No. 9413, National Bureau of Economic Research, 2003.

Klein, Stephen P., Laura S. Hamilton, et al., "What Do Test Scores in Texas Tell Us?" RAND Corporation, 2000.

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

Lindsay, Drew, "Whodunit? Officials Find Thousands of Erasures on Standardized Tests and Suspect Tampering," *Education Week* (1996), 25–29.

Loughran, Regina, and Thomas Comiskey, "Cheating the Children: Educator

Misconduct on Standardized Tests," Report of the City of New York Special Commissioner of Investigation for the New York City School District, 1999.

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

May, Meredith, "State Fears Cheating by Teachers," *San Francisco Chronicle,* October 4, 2000.

Perlman, Carol L., "Results of a Citywide Testing Program Audit in Chicago," Paper for American Educational Research Association annual meeting, Chicago, IL, 1985.

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

Rohlfs, Chris, "Estimating Classroom Learning Using the School Breakfast Program as an Instrument for Attendance and Class Size: Evidence from Chicago Public Schools," Unpublished manuscript, University of Chicago, 2002.

Tysome, Tony, "Cheating Purge: Inspectors out," *New York Times Higher Education Supplement,* August 19, 1994.

Wollack, James A., "A Nominal Response Model Approach for Detecting Answer Copying," *Applied Psychological Measurement,* XX (1997), 144–152.