



*The pressures of accountability
may encourage school personnel
to doctor the results from
high-stakes tests.
Here's how to stop them*

To Catch a Cheat

THE COSTANO SCHOOL IN EAST PALO ALTO, CALIFORNIA, ACHIEVED national recognition in 2000 for overcoming enormous obstacles to achieve academic success. Its principal, Marthelia Hargrove, was designated Principal of the Year by the National Alliance of Black School Educators, and the school was identified by the nonprofit Education Trust as one of the relatively few high-poverty, high-minority schools nationwide that perform well on state exams.

Just two years later, however, the school's performance came under question when 11 former students, interviewed by the *San Jose Mercury News*, acknowledged that they had received inappropriate help from teachers on state tests. A former Costano teacher told the paper that a school administrator had encouraged him to cheat on the exams, and a subsequent examination of the students' actual answer sheets revealed an unusually high number of erasure marks where answers had been changed from wrong to right.

Costano does not appear to be an isolated case. Similar incidents have been documented in Indiana, Maryland, New York, Texas, and Virginia, to name just a few. These scandals have aroused public concern, but there has been little hard evidence on the extent of cheating by school personnel on the type of tests required by recently enacted accountability legislation. Our research on cheating in Chicago indicates that the problem is significant enough to require attention, but not so widespread as to call into question the integrity of the nation's educators. A statistical technique that identifies cases of potential cheating within Chicago's public schools (described in detail below) indicates that, on any given exam, only 3 to 6 percent of classrooms experience instances

ILLUSTRATIONS BY MICHAEL WITTE

by BRIAN A. JACOB & STEVEN D. LEVITT

of teachers or administrators' doctoring students' exams.

Still, with the implementation of the No Child Left Behind Act, the incentives for teachers and administrators to manipulate the results from high-stakes tests will only grow, especially as schools begin to feel the consequences of low scores. No matter what the outcome, cheating hurts the cause of school improvement. When they are not caught, school personnel who cheat may avoid accountability and thereby also cheat their students and the public. When they are caught, these school personnel produce scandals that drain energy and resources from the real business of schools and reform. With better detection methods, like the technique described here, schools can both investigate potential abuses of the system and deter cheating in the future, both of which will increase confidence in the results from high-stakes tests.

This article describes the results of a three-year investigation into cheating by school personnel. The goals of this research were to measure the prevalence of cheating by teachers and administrators and to analyze the factors that predict cheating. Strategies for cheating can include altering students' answer sheets, giving students the answers, or obtaining copies of an exam before the test date and literally "teaching the test"—prepping students with answers to actual test questions. Using data on test scores and student records from the Chicago Public Schools, we developed a statistical algorithm to identify classrooms where cheating was suspected. This method depends on two hallmarks of potential cheating: unexpected fluctuations in students' test scores and unusual patterns of answers for students within a classroom. At the invitation of Arne Duncan, chief education officer of the Chicago Public Schools, we were given the opportunity to work with

Chicago administrators to design and implement auditing and retesting procedures in 2002. The results of this retesting provided strong support for the effectiveness of our method for detecting cheating. Finally, we examined whether cheating responds to incentives, notably the introduction of a high-stakes testing regime in 1996.

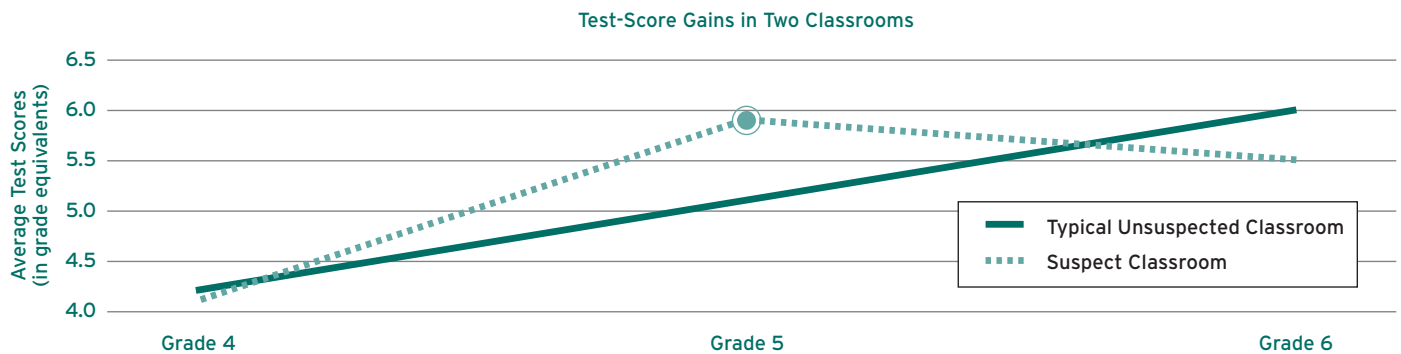
Testing Procedures

Each year, Chicago students in grades 3 through 8 are required to take the Iowa Test of Basic Skills (ITBS), a national norm-referenced exam with a reading comprehension section and three separate math sections, for a total of four subtests. The sample used in our analysis included all students in grades 3–7 for the years 1993–2000. (We excluded 8th graders because our statistical method requires scores for the following year.) Our data included each student's answers on each year's test; which school and classroom each student was in; each student's previous and future test scores; and demographic variables including each student's age, sex, race, and eligibility for the federal school lunch program, a widely used proxy for family income. The data also included a wide range of school-level characteristics. However, the data did not allow us to directly link teachers to classrooms or to track a particular teacher over time. Our final data set contained roughly 20,000 students per grade per year, distributed across approximately 1,000 classrooms, for a total of more than 40,000 "classroom years" of data (with four subject tests per classroom year) and more than 700,000 "student year" observations.

The ITBS is administered over a week-long period in early May. Only 3rd-grade teachers are permitted to give the exam

Suspicious Trend (Figure 1)

In a typical classroom, students' average performance tends to increase by one "grade equivalent" per year. A classroom where cheating may have occurred, by contrast, tends to experience a gain of greater than one grade equivalent, followed the next year by a decline. In the example below, illustrating the trends in two actual Chicago classrooms, the increase in scores from grade 4 to grade 5 with a drop in grade 6 signals a classroom where cheating may be going on.



SOURCE: Authors

to their own students; in other grades, teachers generally switch classrooms to proctor the exam. The exams are 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 designated test coordinator, usually an administrator or counselor. Each section of the exam consists of 30 to 60 multiple-choice questions; students are given between 30 and 75 minutes to complete a section. Students mark their responses on answer sheets, which are scanned electronically to determine the student's score. There is no penalty for guessing.

Once time is up, teachers or administrators collect the exams and then "clean" the answer keys by erasing stray pencil marks, removing dirt or debris from the form, and darkening answers that were marked faintly by students. At the end of the testing week, the testing coordinators at each school deliver the completed answer keys and exams to the Chicago district's central office. School personnel are not supposed to keep copies of the actual exams, but school officials acknowledge that a number of teachers do so each year.

Identifying Cheating by School Personnel

Cheating by teachers and administrators, especially in extreme cases, is likely to leave tell-tale signs. Consider the results from two actual classrooms taking the same test, one of which we suspect of cheating. Figure 1 presents the students' average scores for the preceding, current, and following years, expressed in "grade equivalents." The ITBS is normed in such a way that the average 6th grader, for instance, taking the test during the eighth month of the school year, would score 6.8. A typical student would be expected to gain one grade equivalent for each year of school.

Students in the cheating classroom experienced relatively large increases in test scores from the previous to the current year (1.7 grade equivalents on average). The following year, however, their scores actually *declined*. By comparison, students in



On any given exam, school personnel appear to cheat in only 3 to 6 percent of classrooms.

the typical classroom gained roughly one grade equivalent each year, as expected. The pattern of an unexpectedly large increase and then a decline in test scores fits the profile of a classroom where cheating is going on. Increases in test scores that are the result of cheating do not represent genuine gains in knowledge. Thus there is no reason to expect unusually large

The No Child Left Behind Act will only increase the incentives for teachers and administrators to manipulate the results from high-stakes tests.

increases to be sustained on future exams taken by these students—unless, of course, next year’s teachers also cheat. In the absence of such collusive behavior, large gains due to cheating should be followed by unusually small gains or even declines for these students the next year. If sizable increases in test scores were due to an especially talented teacher, the gains would be likely to have a greater permanent component, even if students regressed a bit the following year.

We developed a measure of how unusual the fluctuations in test scores are by ranking each classroom’s average test-score gains against all other classrooms in that same subject, grade, and year. Classrooms with relatively big gains on this year’s test and relatively small gains on next year’s test will score

high on this indicator. Since students in the same classroom typically disperse into different classrooms the following year, we used the composition of the current year’s classrooms as our basis of analysis.

However, mere fluctuations in test scores are not enough to justify an investigation into potential cheating. Evidence that students’ answers may have been influenced by school personnel in an unethical manner is also necessary. For instance, the 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. In the classroom in Figure 1 with unusual test-score patterns, an examination of students’ exams revealed that a large block of students provided identical answers on consecutive

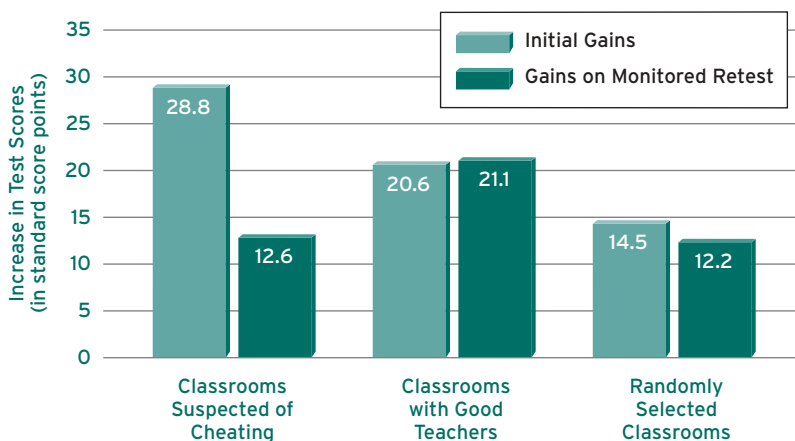
questions in the middle of the test, while no such pattern was found in the other classroom. More sophisticated cheating might involve skipping some questions so as to avoid a large block of identical answers or altering different blocks of questions for different students. We detected these and other potential strategies for cheating by combining four different measures of how suspicious a classroom’s answer strings appear (see the web version of this article at www.educationnext.org for details). The essence of each measure is to predict whether individual students would answer each question correctly based on their past and future performance and to use these predictions to identify unexpected answer patterns.

These two indicators of cheating—unusually large increases followed by small gains or even declines in test scores the next year and unexpected patterns in students’ answers—form the basis of our method for detecting evidence of suspected cheating. Among classrooms that are extremely unlikely to have cheated (classrooms with low values on both cheating indicators), there is only a slight positive relationship between unusual test scores and suspicious answer strings. By contrast, among classrooms with highly suspicious answer strings, the probability of large fluctuations

In Classrooms Suspected of Cheating, Scores Fall on Monitored Retest (Figure 2)

When classrooms suspected of cheating were retested under monitored conditions, their average gains on the Iowa Test of Basic Skills reading exam dropped by more than half. Classrooms that had experienced equally large gains but were not suspected of cheating made similar gains the second time around, a sign of good teaching. Randomly selected classrooms that were retested maintained almost all their gains.

Comparison of Results for Initial Test and Monitored Retest*



* Gains on Iowa Test of Basic Skills between Spring 2001 and Spring 2002.

SOURCE: Authors

in test scores increases dramatically. To estimate the prevalence of cheating, we simply take the difference between the number of classrooms with highly suspicious answer strings that one would predict to have had large test-score fluctuations (assuming that they hadn't cheated) and the actual number of classrooms with highly suspicious answer strings and large test-score fluctuations. Intuitively, this measures the number of "extra" classrooms we find with large test-score fluctuations among the set of classrooms with highly suspicious answer strings.

Our analysis estimated that the share of classrooms where cheating occurred on any particular subject test ranged from 1.1 percent to 2.1 percent, depending on where the statistical cutoff point was set. For instance, the higher estimate has a cutoff at the 80th percentile, meaning that to be suspected of cheating, a classroom would need to experience gains better than 80 percent of all classrooms, and its pattern of answers would need to appear more suspicious than 80 percent of all classrooms. The same is true for the lower estimate, which requires that classrooms be above the 95th percentile on both indicators.

If every classroom that cheated did so on only one subject test, then the overall prevalence of cheating on the ITBS would simply be four times its prevalence on any particular subject test. In many instances, however, classrooms appeared to cheat on multiple subjects. Thus our method estimated that cheating occurred on at least one subject test in 3.4 to 5.6 percent of all classrooms. The range in estimates is again explained by the higher and lower statistical cutoff points we used to determine whether a classroom should be suspected of cheating.

This pattern of results indicates that cheating on one subject test is not an isolated affair. Therefore, if we have detected genuine cheating, 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 cheated 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 school's testing coordinator, one would expect to find multiple classrooms in a school cheating in any given year and perhaps even that cheating in a school one year predicts cheating there in future years.

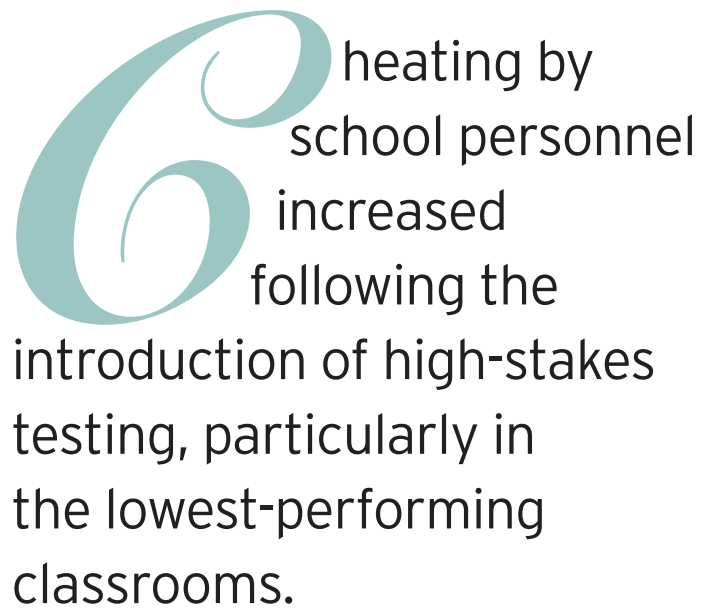
Our analysis indicated that our measures of cheating were strongly correlated across all of these dimensions. For example, if wholesale cheating had occurred on one test in a given subject within a particular classroom, it was ten times more likely that cheating on a test in another subject would be identified in the same classroom. There was also evidence of correlation in cheating within schools, suggesting some centralized effort by a school counselor, test coordinator, or the principal. Finally, we found that cheating in the classroom the previous year predicted cheating this year. Classrooms that cheated in the same subject the previous year were nearly nine times more likely to cheat this year than a

classroom that did not cheat last year.

The Retesting Experiment

Because of the necessarily indirect nature of our empirical strategy, we explored a range of supplemental tests designed to assess the validity of the estimates. These all confirmed that our methods are highly effective at isolating cases of potential cheating and avoiding false positives (innocent teachers accused of cheating). But perhaps the most compelling evidence for the validity of our methods comes from the results of a unique policy intervention.

A few weeks after the spring 2002 ITBS exam was administered, the Chicago Public Schools gave us the opportunity to retest more than 100 classrooms under controlled circumstances that made cheating virtually impossible. The retested classrooms were of three types. First, we identified the set of classrooms most likely to have cheated, based on the prevalence of suspicious answer strings and unusual gains in test scores



cheating by school personnel increased following the introduction of high-stakes testing, particularly in the lowest-performing classrooms.

on the initial spring 2002 exam. The second type of classroom, which we used as a control group, consisted of classrooms with large test-score gains but no evidence of cheating in their answer strings, a sign of plain-old good teaching. These students would be expected to maintain their gains when retested, though there might be some regression. The third type of classroom, which served as another control group, consisted of randomly selected classrooms.

Classrooms identified as "most likely to have cheated" experienced gains on the initial spring 2002 test that were nearly twice as large as the typical Chicago classroom (see Figure 2).

On the retest, however, those excess gains disappeared completely—the gains between the spring 2001 test and the spring 2002 retest were close to the system-wide average. In stark contrast, classrooms identified as potentially having good teachers scored even higher on the reading retest than they did on the initial test. Their math scores fell slightly on the retest, but they continued to post extremely large gains. The randomly selected classrooms also maintained almost all of their gains when retested, as would be expected.



What Predicts Cheating?


Using the indicators of teacher cheating described above, we conducted a series of analyses to examine the relationship between cheating and a variety of classroom and school characteristics. Several striking findings emerged.

On the one hand, classrooms that performed poorly the previous year were much more likely to cheat. For example, a classroom that scored one standard deviation below average the previous year was 23 percent more likely to cheat the next year. Likewise, classrooms in schools with lower achievement, higher poverty rates, and more African-American students were more likely to cheat. Classrooms with a higher proportion of students included in the official test reporting (in other words, fewer kids designated as disabled or limited English proficient and thus excluded from the accountability system) were also more likely to cheat—a 10 percentage point increase in the proportion of students in a class whose test scores “count” increased the likelihood of cheating by roughly 20 percent. Because a greater portion of their students would contribute to the overall assessment of the classroom and school under the accountability policy, these teachers perhaps felt more pressure to ensure that the students in their classrooms scored high on the exams. Teachers who administered the exam to their own students were approximately 50 percent more likely to cheat.

On the other hand, classrooms with students from multiple grades were 65 percent *less* likely to cheat than classrooms where all students were in the same grade. It is probably 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. Classrooms in schools with teachers who graduated from

more-prestigious undergraduate institutions were also less likely to cheat; classrooms in schools with younger teachers were more likely to cheat.

Reusing a test form that had been administered in a previous year had no statistically significant impact on cheating. This suggests that teachers’ taking old exams and teaching specific questions to stu-



heating could be virtually eliminated at a relatively low cost through the implementation of safeguards like those used by the Educational Testing Service.

dents is not an important component of what we are detecting as cheating (though anecdotal evidence suggests that this practice exists).

From a public-policy perspective, especially in the era of high-stakes testing, perhaps the most important question is how cheating responds to incentives. Before 1996, ITBS scores merely provided teachers and parents with a sense of how a child was progressing academically. But beginning in 1996, the Chicago school district launched an initiative designed to hold students and teachers accountable for student learning.

The reform agenda 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 exam. (Students’ performance in math was not used to determine probation status.) Schools placed on probation and exhibiting little subsequent improvement could be reconstituted; this meant closing the school and dismissing or reassigning all personnel. 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 performance. Under the new policy, students in the 3rd, 6th, and 8th grades had to meet minimum standards on the ITBS in both reading and mathematics in order to be promoted to the next grade. The promotion standards were implemented in the spring of 1997 for 3rd and 6th graders. Decisions on promotion were based solely on students’ scores in reading comprehension and mathematics.

As might be expected, the cheating by school personnel increased following the introduction of high-stakes testing, particularly in the lowest-performing classrooms. For example, the likelihood of cheating in a classroom that was one standard deviation below the mean increased by roughly 29 percent in response to the school probation policy and 43 percent due to the ending of social promotion. The magnitude of these changes is particularly great considering that no elementary school on probation was actually reconstituted during this period and that the social promotion policy has no direct effects on teachers’ pay or job security. By contrast, classrooms that performed one standard deviation above the average experienced no significant change in cheating in response to these two policies.

Conclusions

The results of this study demonstrate the value of statistical analysis to school districts interested in catching cheaters or deterring future cheating. The findings from the retesting experiment were used to launch investigations of 29 classrooms. While these investigations have not been completed, it is expected that disciplinary action will be brought against a substantial number of teachers, test administrators, and principals. Perhaps more important, a preliminary analysis of the 2003 test results suggests that the incidence of cheating has declined in the district.

Moreover, there is a more positive aspect to our methods than just the isolation of instances of potential cheating. Using these tools, we were able to identify a set of classrooms that made extraordinary test-score gains without any indication of cheating. This will pave the way to identifying and rewarding outstanding teachers.

While evidence of cheating is sometimes used to impugn high-stakes testing programs, our results actually show that explicit cheating by school personnel is not likely to be a serious enough problem by itself to call into question high-stakes testing, both because the most egregious forms of cheating are relatively rare and, more important, because cheating could be virtually eliminated at a relatively low cost through the implementation of proper safeguards, such as those used by the Educational Testing Service on the SAT or GRE exams.

However, the sort of cheating that our methods are capable of catching is just one of many potential behavioral responses to high-stakes testing. Other responses, like teaching to the test and cheating in a subtler manner, such as giving the students extra time, are presumably also present but harder to measure. The challenge for educators and policymakers will be to develop a system that captures the obvious benefits of high-stakes testing as a means of providing incentives while minimizing the possible distortions that these measures induce.

—Brian A. Jacob is an assistant professor at Harvard University’s Kennedy School of Government, and Steven D. Levitt is a professor of economics and social sciences at the University of Chicago. A more detailed account of this investigation can be found in the August 2003 issue of the *Quarterly Journal of Economics*.