1 What Do Schoolteachers and Sumo Wrestlers Have in Common?

Imagine for a moment that you are the manager of a day-care center. You have a clearly stated policy that children are supposed to be picked up by 4 p.m. But very often parents are late. The result: at day's end, you have some anxious children and at least one teacher who must wait around for the parents to arrive. What to do?

A pair of economists who heard of this dilemma—it turned out to be a rather common one—offered a solution: fine the tardy parents. Why, after all, should the day-care center take care of these kids for free?

The economists decided to test their solution by conducting a study of ten day-care centers in Haifa, Israel. The study lasted twenty weeks, but the fine was not introduced immediately. For the first four weeks, the economists simply kept track of the number of parents who came late; there were, on average, eight late pickups per week per day-care center. In the fifth week, the fine was enacted. It was announced that any parent arriving more than ten minutes late

would pay $3 per child for each incident. The fee would be added to the parents' monthly bill, which was roughly $380.

After the fine was enacted, the number of late pickups promptly went . . . up. Before long there were twenty late pickups per week, more than double the original average. The incentive had plainly backfired.

Economics is, at root, the study of incentives: how people get what they want, or need, especially when other people want or need the same thing. Economists love incentives. They love to dream them up and enact them, study them and tinker with them. The typical economist believes the world has not yet invented a problem that he cannot fix if given a free hand to design the proper incentive scheme. His solution may not always be pretty—it may involve coercion or exorbitant penalties or the violation of civil liberties—but the original problem, rest assured, will be fixed. An incentive is a bullet, a lever, a key: an often tiny object with astonishing power to change a situation.

We all learn to respond to incentives, negative and positive, from the outset of life. If you toddle over to the hot stove and touch it, you burn a finger. But if you bring home straight A's from school, you get a new bike. If you are spotted picking your nose in class, you get ridiculed. But if you make the basketball team, you move up the social ladder. If you break curfew, you get grounded. But if you ace your SATs, you get to go to a good college. If you flunk out of law school, you have to go to work at your father's insurance company. But if you perform so well that a rival company comes calling, you become a vice president and no longer have to work for your father. If you become so excited about your new vice president job that you drive home at eighty mph, you get pulled over by the police and fined $100. But if you hit your sales projections and collect a year-end bonus, you not only aren't worried about the $100 ticket but can also afford to buy

that Viking range you've always wanted—and on which your toddler can now burn her own finger.

An incentive is simply a means of urging people to do more of a good thing and less of a bad thing. But most incentives don't come about organically. Someone—an economist or a politician or a parent—has to invent them. Your three-year-old eats all her vegetables for a week? She wins a trip to the toy store. A big steelmaker belches too much smoke into the air? The company is fined for each cubic foot of pollutants over the legal limit. Too many Americans aren't paying their share of income tax? It was the economist Milton Friedman who helped come up with a solution to this one: automatic tax withholding from employees' paychecks.

There are three basic flavors of incentive: economic, social, and moral. Very often a single incentive scheme will include all three varieties. Think about the anti-smoking campaign of recent years. The addition of a $3-per-pack "sin tax" is a strong economic incentive against buying cigarettes. The banning of cigarettes in restaurants and bars is a powerful social incentive. And when the U.S. government asserts that terrorists raise money by selling black-market cigarettes, that acts as a rather jarring moral incentive.

Some of the most compelling incentives yet invented have been put in place to deter crime. Considering this fact, it might be worthwhile to take a familiar question—why is there so much crime in modern society?—and stand it on its head: why isn't there a lot *more* crime?

After all, every one of us regularly passes up opportunities to maim, steal, and defraud. The chance of going to jail—thereby losing your job, your house, and your freedom, all of which are essentially economic penalties—is certainly a strong incentive. But when it comes to crime, people also respond to moral incentives (they don't want to do something they consider wrong) and social incentives

(they don't want to be seen by others as doing something wrong). For certain types of misbehavior, social incentives are terribly powerful. In an echo of Hester Prynne's scarlet letter, many American cities now fight prostitution with a "shaming" offensive, posting pictures of convicted johns (and prostitutes) on websites or on local-access television. Which is a more horrifying deterrent: a $500 fine for soliciting a prostitute or the thought of your friends and family ogling you on www.HookersAndJohns.com.

So through a complicated, haphazard, and constantly readjusted web of economic, social, and moral incentives, modern society does its best to militate against crime. Some people would argue that we don't do a very good job. But taking the long view, that is clearly not true. Consider the historical trend in homicide (not including wars), which is both the most reliably measured crime and the best barometer of a society's overall crime rate. These statistics, compiled by the criminologist Manuel Eisner, track the historical homicide levels in five European regions.

HOMICIDES

(per 100,000 People)

	England	Netherlands and Belgium	Scandinavia	Germany and Switzerland	Italy
13th and 14th c.	23.0	47.0	n.a.	37.0	56.0
15th c.	n.a.	45.0	46.0	16.0	73.0
16th c.	7.0	25.0	21.0	11.0	47.0
17th c.	5.0	7.5	18.0	7.0	32.0
18th c.	1.5	5.5	1.9	7.5	10.5
19th c.	1.7	1.6	1.1	2.8	12.6
1900–1949	0.8	1.5	0.7	1.7	3.2
1950–1994	0.9	0.9	0.9	1.0	1.5

The steep decline of these numbers over the centuries suggests that, for one of the gravest human concerns—getting murdered—the incentives that we collectively cook up are working better and better.

So what was wrong with the incentive at the Israeli day-care centers?

You have probably already guessed that the $3 fine was simply too small. For that price, a parent with one child could afford to be late every day and only pay an extra $60 each month—just one-sixth of the base fee. As babysitting goes, that's pretty cheap. What if the fine had been set at $100 instead of $3? That would have likely put an end to the late pickups, though it would have also engendered plenty of ill will. (Any incentive is inherently a trade-off; the trick is to balance the extremes.)

But there was another problem with the day-care center fine. It substituted an economic incentive (the $3 penalty) for a moral incentive (the guilt that parents were supposed to feel when they came late). For just a few dollars each day, parents could buy off their guilt. Furthermore, the small size of the fine sent a signal to the parents that late pickups weren't such a big problem. If the day-care center suffers only $3 worth of pain for each late pickup, why bother to cut short the tennis game? Indeed, when the economists eliminated the $3 fine in the seventeenth week of their study, the number of late-arriving parents didn't change. Now they could arrive late, pay no fine, *and* feel no guilt.

Such is the strange and powerful nature of incentives. A slight tweak can produce drastic and often unforeseen results. Thomas Jefferson noted this while reflecting on the tiny incentive that led to the Boston Tea Party and, in turn, the American Revolution: "So inscrutable is the arrangement of causes and consequences in this world that a two-penny duty on tea, unjustly imposed in a sequestered part of it, changes the condition of all its inhabitants."

In the 1970s, researchers conducted a study that, like the Israeli day-care study, pitted a moral incentive against an economic incentive. In this case, they wanted to learn about the motivation behind blood donations. Their discovery: when people are given a small stipend for donating blood rather than simply being praised for their altruism, they tend to donate *less* blood. The stipend turned a noble act of charity into a painful way to make a few dollars, and it wasn't worth it.

What if the blood donors had been offered an incentive of $50, or $500, or $5,000? Surely the number of donors would have changed dramatically.

But something else would have changed dramatically as well, for every incentive has its dark side. If a pint of blood were suddenly worth $5,000, you can be sure that plenty of people would take note. They might literally steal blood at knifepoint. They might pass off pig blood as their own. They might circumvent donation limits by using fake IDs. Whatever the incentive, whatever the situation, dishonest people will try to gain an advantage by whatever means necessary.

Or, as W. C. Fields once said: a thing worth having is a thing worth cheating for.

Who cheats?

Well, just about anyone, if the stakes are right. You might say to yourself, *I* don't cheat, regardless of the stakes. And then you might remember the time you cheated on, say, a board game. Last week. Or the golf ball you nudged out of its bad lie. Or the time you really wanted a bagel in the office break room but couldn't come up with the dollar you were supposed to drop in the coffee can. And then took the bagel anyway. And told yourself you'd pay double the next time. And didn't.

For every clever person who goes to the trouble of creating an in-

centive scheme, there is an army of people, clever and otherwise, who will inevitably spend even more time trying to beat it. Cheating may or may not be human nature, but it is certainly a prominent feature in just about every human endeavor. Cheating is a primordial economic act: getting more for less. So it isn't just the boldface names—inside-trading CEOs and pill-popping ballplayers and perk-abusing politicians—who cheat. It is the waitress who pockets her tips instead of pooling them. It is the Wal-Mart payroll manager who goes into the computer and shaves his employees' hours to make his own performance look better. It is the third grader who, worried about not making it to the fourth grade, copies test answers from the kid sitting next to him.

Some cheating leaves barely a shadow of evidence. In other cases, the evidence is massive. Consider what happened one spring evening at midnight in 1987: seven million American children suddenly disappeared. The worst kidnapping wave in history? Hardly. It was the night of April 15, and the Internal Revenue Service had just changed a rule. Instead of merely listing each dependent child, tax filers were now required to provide a Social Security number for each child. Suddenly, seven million children—children who had existed only as phantom exemptions on the previous year's 1040 forms—vanished, representing about one in ten of all dependent children in the United States.

The incentive for those cheating taxpayers was quite clear. The same for the waitress, the payroll manager, and the third grader. But what about that third grader's *teacher*? Might she have an incentive to cheat? And if so, how would she do it?

Imagine now that instead of running a day-care center in Haifa, you are running the Chicago Public Schools, a system that educates 400,000 students each year.

The most volatile current debate among American school administrators, teachers, parents, and students concerns "high-stakes" testing. The stakes are considered high because instead of simply testing students to measure their progress, schools are increasingly held accountable for the results.

The federal government mandated high-stakes testing as part of the No Child Left Behind law, signed by President Bush in 2002. But even before that law, most states gave annual standardized tests to students in elementary and secondary school. Twenty states rewarded individual schools for good test scores or dramatic improvement; thirty-two states sanctioned the schools that didn't do well.

The Chicago Public School system embraced high-stakes testing in 1996. Under the new policy, a school with low reading scores would be placed on probation and face the threat of being shut down, its staff to be dismissed or reassigned. The CPS also did away with what is known as social promotion. In the past, only a dramatically inept or difficult student was held back a grade. Now, in order to be promoted, every student in third, sixth, and eighth grade had to manage a minimum score on the standardized, multiple-choice exam known as the Iowa Test of Basic Skills.

Advocates of high-stakes testing argue that it raises the standards of learning and gives students more incentive to study. Also, if the test prevents poor students from advancing without merit, they won't clog up the higher grades and slow down good students. Opponents, meanwhile, worry that certain students will be unfairly penalized if they don't happen to test well, and that teachers may concentrate on the test topics at the exclusion of more important lessons.

Schoolchildren, of course, have had incentive to cheat for as long as there have been tests. But high-stakes testing has so radically changed the incentives for teachers that they too now have added reason to cheat. With high-stakes testing, a teacher whose students test

poorly can be censured or passed over for a raise or promotion. If the entire school does poorly, federal funding can be withheld; if the school is put on probation, the teacher stands to be fired. High-stakes testing also presents teachers with some positive incentives. If her students do well enough, she might find herself praised, promoted, and even richer: the state of California at one point introduced bonuses of $25,000 for teachers who produced big test-score gains.

And if a teacher were to survey this newly incentivized landscape and consider somehow inflating her students' scores, she just might be persuaded by one final incentive: teacher cheating is rarely looked for, hardly ever detected, and just about never punished.

How might a teacher go about cheating? There are any number of possibilities, from the brazen to the sophisticated. A fifth-grade student in Oakland recently came home from school and gaily told her mother that her super-nice teacher had written the answers to the state exam right there on the chalkboard. Such instances are certainly rare, for placing your fate in the hands of thirty prepubescent witnesses doesn't seem like a risk that even the worst teacher would take. (The Oakland teacher was duly fired.) There are more subtle ways to inflate students' scores. A teacher can simply give students extra time to complete the test. If she obtains a copy of the exam early—that is, illegitimately—she can prepare them for specific questions. More broadly, she can "teach to the test," basing her lesson plans on questions from past years' exams, which isn't considered cheating but certainly violates the spirit of the test. Since these tests all have multiple-choice answers, with no penalty for wrong guesses, a teacher might instruct her students to randomly fill in every blank as the clock is winding down, perhaps inserting a long string of Bs or an alternating pattern of Bs and Cs. She might even fill in the blanks for them after they've left the room.

But if a teacher *really* wanted to cheat—and make it worth her

while—she might collect her students' answer sheets and, in the hour or so before turning them in to be read by an electronic scanner, erase the wrong answers and fill in correct ones. (And you always thought that no. 2 pencil was for the *children* to change their answers.) If this kind of teacher cheating is truly going on, how might it be detected?

To catch a cheater, it helps to think like one. If you were willing to erase your students' wrong answers and fill in correct ones, you probably wouldn't want to change too many wrong answers. That would clearly be a tip-off. You probably wouldn't even want to change answers on every student's test—another tip-off. Nor, in all likelihood, would you have enough time, because the answer sheets are turned in soon after the test is over. So what you might do is select a string of eight or ten consecutive questions and fill in the correct answers for, say, one-half or two-thirds of your students. You could easily memorize a short pattern of correct answers, and it would be a lot faster to erase and change that pattern than to go through each student's answer sheet individually. You might even think to focus your activity toward the end of the test, where the questions tend to be harder than the earlier questions. In that way, you'd be most likely to substitute correct answers for wrong ones.

If economics is a science primarily concerned with incentives, it is also—fortunately—a science with statistical tools to measure how people respond to those incentives. All you need are some data.

In this case, the Chicago Public School system obliged. It made available a database of the test answers for every CPS student from third grade through seventh grade from 1993 to 2000. This amounts to roughly 30,000 students per grade per year, more than 700,000 sets of test answers, and nearly 100 million individual answers. The data, organized by classroom, included each student's question-by-question answer strings for reading and math tests. (The actual paper answer sheets were not included; they were habitually shredded soon

after a test.) The data also included some information about each teacher and demographic information for every student, as well as his or her past and future test scores—which would prove a key element in detecting the teacher cheating.

Now it was time to construct an algorithm that could tease some conclusions from this mass of data. What might a cheating teacher's classroom look like?

The first thing to search for would be unusual answer patterns in a given classroom: blocks of identical answers, for instance, especially among the harder questions. If ten very bright students (as indicated by past and future test scores) gave correct answers to the exam's first five questions (typically the easiest ones), such an identical block shouldn't be considered suspicious. But if ten poor students gave correct answers to the *last* five questions on the exam (the hardest ones), that's worth looking into. Another red flag would be a strange pattern within any one student's exam—such as getting the hard questions right while missing the easy ones—especially when measured against the thousands of students in other classrooms who scored similarly on the same test. Furthermore, the algorithm would seek out a classroom full of students who performed far better than their past scores would have predicted and who then went on to score significantly lower the following year. A dramatic one-year spike in test scores might initially be attributed to a *good* teacher; but with a dramatic fall to follow, there's a strong likelihood that the spike was brought about by artificial means.

Consider now the answer strings from the students in two sixth-grade Chicago classrooms who took the identical math test. Each horizontal row represents one student's answers. The letter a, b, c, or d indicates a correct answer; a number indicates a wrong answer, with 1 corresponding to a, 2 corresponding to b, and so on. A zero represents an answer that was left blank. One of these classrooms almost cer-

tainly had a cheating teacher and the other did not. Try to tell the difference—although be forewarned that it's not easy with the naked eye.

Classroom A

112a4a342cb214d0001acd24a3a12dadbcb4a0000000
d4a2341cacbddad3142a2344a2ac23421c00adb4b3cb
1b2a34d4ac42d23b141acd24a3a12dadbcb4a2134141
dbaab3dcacb1dadbc42ac2cc31012dadbcb4adb40000
d12443d43232d32323c213c22d2c23234c332db4b300
db2abad1acbdda212b1acd24a3a12dadbcb400000000
d4aab2124cbddadbcb1a42cca3412dadbcb423134bc1
1b33b4d4a2b1dadbc3ca22c0000000000000000000
d43a3a24acb1d32b412acd24a3a12dadbcb422143bc0
313a3ad1ac3d2a23431223c000012dadbcb400000000
db2a33dcacbd32d313c21142323cc300000000000000
d43ab4d1ac3dd43421240d24a3a12dadbcb400000000
db223a24acb11a3b24cacd12a241cdadbcb4adb4b300
db4abadcacb1dad3141ac212a3a1c3a144ba2db41b43
1142340c2cbddadb4b1acd24a3a12dadbcb43d133bc4
214ab4dc4cbdd31b1b2213c4ad412dadbcb4adb00000
1423b4d4a23d24131413234123a243a2413a21441343
3b3ab4d14c3d2ad4cbcac1c003a12dadbcb4adb40000
dba2ba21ac3d2ad3c4c4cd40a3a12dadbcb400000000
d122ba2cacbd1a13211a2d02a2412d0dbcb4adb4b3c0
144a3adc4cbddadbcbc2c2cc43a12dadbcb4211ab343
d43aba3cacbddadbcbca42c2a3212dadbcb42344b3cb

Classroom B

db3a431422bd131b4413cd422a1acda332342d3ab4c4
d1aa1a11acb2d3dbc1ca22c23242c3a142b3adb243c1

d42a12d2a4b1d32b21ca2312a3411d00000000000000
3b2a34344c32d21b1123cdc0000000000000000000000
34aabad12cbdd3d4c1ca112cad2ccd00000000000000
d33a3431a2b2d2d44b2acd2cad2c2223b40000000000
23aa32d2a1bd2431141342c13d212d233c34a3b3b000
d32234d4a1bdd23b242a22c2a1a1cda2b1baa33a0000
d3aab23c4cbddadb23c322c2a222223232b443b24bc3
d13a14313c31d42b14c421c42332cd2242b3433a3343
d13a3ad122b1da2b11242dc1a3a12100000000000000
d12a3ad1a13d23d3cb2a21ccada24d2131b440000000
314a133c4cbd142141ca424cad34c122413223ba4b40
d42a3adcacbddadbc42ac2c2ada2cda341baa3b24321
db1134dc2cb2dadb24c412c1ada2c3a341ba20000000
d1341431acbddad3c4c213412da22d3d1132a1344b1b
1ba41a21a1b2dadb24ca22c1ada2cd32413200000000
dbaa33d2a2bddadbcbca11c2a2accda1b2ba20000000

If you guessed that classroom A was the cheating classroom, congratulations. Here again are the answer strings from classroom A, now reordered by a computer that has been asked to apply the cheating algorithm and seek out suspicious patterns.

Classroom A

(With cheating algorithm applied)

1. 112a4a342cb214d0001**acd24a3a12dadbcb4**a0000000
2. 1b2a34d4ac42d23b141**acd24a3a12dadbcb4**a2134141
3. db2abad1acbdda212b1**acd24a3a12dadbcb4**00000000
4. d43a3a24acb1d32b412**acd24a3a12dadbcb4**22143bc0
5. d43ab4d1ac3dd43421240d24**a3a12dadbcb4**00000000
6. 1142340c2cbddadb4b1acd24**a3a12dadbcb4**3d133bc4
7. dba2ba21ac3d2ad3c4c4cd40**a3a12dadbcb4**00000000

8. 144a3adc4cbddadbcbc2c2cc4**3a12dadbcb4**211ab343
9. 3b3ab4d14c3d2ad4cbcac1c00**3a12dadbcb4**adb40000
10. d43aba3cacbddadbcbca42c2a32**12dadbcb4**2344b3cb
11. 214ab4dc4cbdd31b1b2213c4ad4**12dadbcb4**adb00000
12. 313a3ad1ac3d2a23431223c0000**12dadbcb4**00000000
13. d4aab2124cbddadbcb1a42cca34**12dadbcb4**23134bc1
14. dbaab3dcacb1dadbc42ac2cc310**12dadbcb4**adb40000
15. db223a24acb11a3b24cacd12a241c**dadbcb4**adb4b300
16. d122ba2cacbd1a13211a2d02a2412d0dbcb4adb4b3c0
17. 1423b4d4a23d24131413234123a243a2413a21441343
18. db4abadcacb1dad3141ac212a3a1c3a144ba2db41b43
19. db2a33dcacbd32d313c21142323cc300000000000000
20. 1b33b4d4a2b1dadbc3ca22c0000000000000000000000
21. d12443d43232d32323c213c22d2c23234c332db4b300
22. d4a2341cacbddad3142a2344a2ac23421c00adb4b3cb

Take a look at the answers in bold. Did fifteen out of twenty-two students somehow manage to reel off the same six consecutive correct answers (the d-a-d-b-c-b string) all by themselves?

There are at least four reasons this is unlikely. One: those questions, coming near the end of the test, were harder than the earlier questions. Two: these were mainly subpar students to begin with, few of whom got six consecutive right answers elsewhere on the test, making it all the more unlikely they would get right the same six hard questions. Three: up to this point in the test, the fifteen students' answers were virtually uncorrelated. Four: three of the students (numbers 1, 9, and 12) left at least one answer blank *before* the suspicious string and then ended the test with another string of blanks. This suggests that a long, unbroken string of blank answers was broken not by the student but by the teacher.

There is another oddity about the suspicious answer string. On

nine of the fifteen tests, the six correct answers are preceded by another identical string, 3-a-1-2, which includes three of four *incorrect* answers. And on all fifteen tests, the six correct answers are followed by the same incorrect answer, a 4. Why on earth would a cheating teacher go to the trouble of erasing a student's test sheet and then fill in the *wrong* answer?

Perhaps she is merely being strategic. In case she is caught and hauled into the principal's office, she could point to the wrong answers as proof that she didn't cheat. Or perhaps—and this is a less charitable but just as likely answer—she doesn't know the right answers herself. (With standardized tests, the teacher is typically not given an answer key.) If this is the case, then we have a pretty good clue as to why her students are in need of inflated grades in the first place: they have a bad teacher.

Another indication of teacher cheating in classroom A is the class's overall performance. As sixth graders who were taking the test in the eighth month of the academic year, these students needed to achieve an average score of 6.8 to be considered up to national standards. (Fifth graders taking the test in the eighth month of the year needed to score 5.8, seventh graders 7.8, and so on.) The students in classroom A averaged 5.8 on their sixth-grade tests, which is a full grade level below where they should be. So plainly these are poor students. A year earlier, however, these students did even worse, averaging just 4.1 on their fifth-grade tests. Instead of improving by one full point between fifth and sixth grade, as would be expected, they improved by 1.7 points, nearly two grades' worth. But this miraculous improvement was short-lived. When these sixth-grade students reached seventh grade, they averaged 5.5—more than two grade levels below standard and even *worse* than they did in sixth grade. Consider the erratic year-to-year scores of three particular students from classroom A:

	5th Grade Score	6th Grade Score	7th Grade Score
Student 3	3.0	6.5	5.1
Student 6	3.6	6.3	4.9
Student 14	3.8	7.1	5.6

The three-year scores from classroom B, meanwhile, are also poor but at least indicate an honest effort: 4.2, 5.1, and 6.0. So an entire roomful of children in classroom A suddenly got very smart one year and very dim the next, or more likely, their sixth-grade teacher worked some magic with a no. 2 pencil.

There are two noteworthy points to be made about the children in classroom A, tangential to the cheating itself. The first is that they are obviously in terrible academic shape, which makes them the very children whom high-stakes testing is promoted as helping the most. The second point is that these students would be in for a terrible shock once they reached the seventh grade. All they knew was that they had been successfully promoted due to their test scores. (No child left behind, indeed.) *They* weren't the ones who artificially jacked up their scores; they probably expected to do great in the seventh grade—and then they failed miserably. This may be the cruelest twist yet in high-stakes testing. A cheating teacher may tell herself that she is helping her students, but the fact is that she would appear far more concerned with helping herself.

An analysis of the entire Chicago data reveals evidence of teacher cheating in more than two hundred classrooms per year, roughly 5 percent of the total. This is a conservative estimate, since the algorithm was able to identify only the most egregious form of cheating—in which teachers systematically changed students' answers—and not the many subtler ways a teacher might cheat. In a recent study among North Carolina schoolteachers, some 35 percent of the respondents said they had witnessed their colleagues cheating in some fashion,

whether by giving students extra time, suggesting answers, or manually changing students' answers.

What are the characteristics of a cheating teacher? The Chicago data show that male and female teachers are about equally prone to cheating. A cheating teacher tends to be younger and less qualified than average. She is also more likely to cheat after her incentives change. Because the Chicago data ran from 1993 to 2000, it bracketed the introduction of high-stakes testing in 1996. Sure enough, there was a pronounced spike in cheating in 1996. Nor was the cheating random. It was the teachers in the lowest-scoring classrooms who were most likely to cheat. It should also be noted that the $25,000 bonus for California teachers was eventually revoked, in part because of suspicions that too much of the money was going to cheaters.

Not every result of the Chicago cheating analysis was so dour. In addition to detecting cheaters, the algorithm could also identify the best teachers in the school system. A good teacher's impact was nearly as distinctive as a cheater's. Instead of getting random answers correct, her students would show real improvement on the easier types of questions they had previously missed, an indication of actual learning. And a good teacher's students carried over all their gains into the next grade.

Most academic analyses of this sort tend to languish, unread, on a dusty library shelf. But in early 2002, the new CEO of the Chicago Public Schools, Arne Duncan, contacted the study's authors. He didn't want to protest or hush up their findings. Rather, he wanted to make sure that the teachers identified by the algorithm as cheaters were truly cheating—and then do something about it.

Duncan was an unlikely candidate to hold such a powerful job. He was only thirty-six when appointed, a onetime academic all-American at Harvard who later played pro basketball in Australia. He

had spent just three years with the CPS—and never in a job important enough to have his own secretary—before becoming its CEO. It didn't hurt that Duncan had grown up in Chicago. His father taught psychology at the University of Chicago; his mother ran an afterschool program for forty years, without pay, in a poor neighborhood. When Duncan was a boy, his afterschool playmates were the underprivileged kids his mother cared for. So when he took over the public schools, his allegiance lay more with schoolchildren and their families than with teachers and their union.

The best way to get rid of cheating teachers, Duncan had decided, was to readminister the standardized exam. He only had the resources to retest 120 classrooms, however, so he asked the creators of the cheating algorithm to help choose which classrooms to test.

How could those 120 retests be used most effectively? It might have seemed sensible to retest only the classrooms that likely had a cheating teacher. But even if their retest scores were lower, the teachers could argue that the students did worse merely because they were told that the scores wouldn't count in their official record—which, in fact, all retested students would be told. To make the retest results convincing, some non-cheaters were needed as a control group. The best control group? The classrooms shown by the algorithm to have the best teachers, in which big gains were thought to have been legitimately attained. If those classrooms held their gains while the classrooms with a suspected cheater lost ground, the cheating teachers could hardly argue that their students did worse only because the scores wouldn't count.

So a blend was settled upon. More than half of the 120 retested classrooms were those suspected of having a cheating teacher. The remainder were divided between the supposedly excellent teachers (high scores but no suspicious answer patterns) and, as a further control, classrooms with mediocre scores and no suspicious answers.

The retest was given a few weeks after the original exam. The chil-

dren were not told the reason for the retest. Neither were the teachers. But they may have gotten the idea when it was announced that CPS officials, not the teachers, would administer the test. The teachers were asked to stay in the classroom with their students, but they would not be allowed to even touch the answer sheets.

The results were as compelling as the cheating algorithm had predicted. In the classrooms chosen as controls, where no cheating was suspected, scores stayed about the same or even rose. In contrast, the students with the teachers identified as cheaters scored far worse, by an average of more than a full grade level.

As a result, the Chicago Public School system began to fire its cheating teachers. The evidence was only strong enough to get rid of a dozen of them, but the many other cheaters had been duly warned. The final outcome of the Chicago study is further testament to the power of incentives: the following year, cheating by teachers fell more than 30 percent.

You might think that the sophistication of teachers who cheat would increase along with the level of schooling. But an exam given at the University of Georgia in the fall of 2001 disputes that idea. The course was called Coaching Principles and Strategies of Basketball, and the final grade was based on a single exam that had twenty questions. Among the questions:

How many halves are in a college basketball game?

a. 1 b. 2 c. 3 d. 4

How many points does a 3-pt. field goal account for in a basketball game?

a. 1 b. 2 c. 3 d. 4

What is the name of the exam which all high school seniors in the State of Georgia must pass?

a. Eye Exam
b. How Do the Grits Taste Exam
c. Bug Control Exam
d. Georgia Exit Exam

In your opinion, who is the best Division I assistant coach in the country?

a. Ron Jirsa
b. John Pelphrey
c. Jim Harrick Jr.
d. Steve Wojciechowski

If you are stumped by the final question, it might help to know that Coaching Principles was taught by Jim Harrick Jr., an assistant coach with the university's basketball team. It might also help to know that his father, Jim Harrick Sr., was the head basketball coach. Not surprisingly, Coaching Principles was a favorite course among players on the Harricks' team. Every student in the class received an A. Not long afterward, both Harricks were relieved of their coaching duties.

If it strikes you as disgraceful that Chicago schoolteachers and University of Georgia professors will cheat—a teacher, after all, is meant to instill values along with the facts—then the thought of cheating among sumo wrestlers may also be deeply disturbing. In Japan, sumo is not only the national sport but also a repository of the country's religious, military, and historical emotion. With its purification rituals and its imperial roots, sumo is sacrosanct in a way that American

sports can never be. Indeed, sumo is said to be less about competition than about honor itself.

It is true that sports and cheating go hand in hand. That's because cheating is more common in the face of a bright-line incentive (the line between winning and losing, for instance) than with a murky incentive. Olympic sprinters and weightlifters, cyclists in the Tour de France, football linemen and baseball sluggers: they have all been shown to swallow whatever pill or powder may give them an edge. It is not only the participants who cheat. Cagey baseball managers try to steal an opponent's signs. In the 2002 Winter Olympic figure-skating competition, a French judge and a Russian judge were caught trying to swap votes to make sure their skaters medaled. (The man accused of orchestrating the vote swap, a reputed Russian mob boss named Alimzhan Tokhtakhounov, was also suspected of rigging beauty pageants in Moscow.)

An athlete who gets caught cheating is generally condemned, but most fans at least appreciate his motive: he wanted so badly to win that he bent the rules. (As the baseball player Mark Grace once said, "If you're not cheating, you're not trying.") An athlete who cheats to *lose*, meanwhile, is consigned to a deep circle of sporting hell. The 1919 Chicago White Sox, who conspired with gamblers to throw the World Series (and are therefore known forever as the Black Sox), retain a stench of iniquity among even casual baseball fans. The City College of New York's championship basketball team, once beloved for its smart and scrappy play, was instantly reviled when it was discovered in 1951 that several players had taken mob money to shave points—intentionally missing baskets to help gamblers beat the point spread. Remember Terry Malloy, the tormented former boxer played by Marlon Brando in *On the Waterfront*? As Malloy saw it, all his troubles stemmed from the one fight in which he took a dive. Otherwise, he could have had class; he could have been a contender.

If cheating to lose is sport's premier sin, and if sumo wrestling is

the premier sport of a great nation, cheating to lose couldn't possibly exist in sumo. Could it?

Once again, the data can tell the story. As with the Chicago school tests, the data set under consideration here is surpassingly large: the results from nearly every official sumo match among the top rank of Japanese sumo wrestlers between January 1989 and January 2000, a total of 32,000 bouts fought by 281 different wrestlers.

The incentive scheme that rules sumo is intricate and extraordinarily powerful. Each wrestler maintains a ranking that affects every slice of his life: how much money he makes, how large an entourage he carries, how much he gets to eat, sleep, and otherwise take advantage of his success. The sixty-six highest-ranked wrestlers in Japan, comprising the *makuuchi* and *juryo* divisions, make up the sumo elite. A wrestler near the top of this elite pyramid may earn millions and is treated like royalty. Any wrestler in the top forty earns at least $170,000 a year. The seventieth-ranked wrestler in Japan, meanwhile, earns only $15,000 a year. Life isn't very sweet outside the elite. Low-ranked wrestlers must tend to their superiors, preparing their meals and cleaning their quarters and even soaping up their hardest-to-reach body parts. So ranking is everything.

A wrestler's ranking is based on his performance in the elite tournaments that are held six times a year. Each wrestler has fifteen bouts per tournament, one per day over fifteen consecutive days. If he finishes the tournament with a winning record (eight victories or better), his ranking will rise. If he has a losing record, his ranking falls. If it falls far enough, he is booted from the elite rank entirely. The eighth victory in any tournament is therefore critical, the difference between promotion and demotion; it is roughly four times as valuable in the rankings as the typical victory.

So a wrestler entering the final day of a tournament on the bubble, with a 7–7 record, has far more to gain from a victory than an opponent with a record of 8–6 has to lose.

Is it possible, then, that an 8–6 wrestler might allow a 7–7 wrestler to beat him? A sumo bout is a concentrated flurry of force and speed and leverage, often lasting only a few seconds. It wouldn't be very hard to let yourself be tossed. Let's imagine for a moment that sumo wrestling *is* rigged. How might we measure the data to prove it?

The first step would be to isolate the bouts in question: those fought on a tournament's final day between a wrestler on the bubble and a wrestler who has already secured his eighth win. (Because more than half of all wrestlers end a tournament with either seven, eight, or nine victories, hundreds of bouts fit these criteria.) A final-day match between two 7–7 wrestlers isn't likely to be fixed, since both fighters badly need the victory. A wrestler with ten or more victories probably wouldn't throw a match either, since he has his own strong incentive to win: the $100,000 prize for overall tournament champion and a series of $20,000 prizes for the "outstanding technique" award, "fighting spirit" award, and others.

Let's now consider the following statistic, which represents the hundreds of matches in which a 7–7 wrestler faced an 8–6 wrestler on a tournament's final day. The left column tallies the probability, based on all past meetings between the two wrestlers fighting that day, that the 7–7 wrestler will win. The right column shows how often the 7–7 wrestler actually did win.

7-7 Wrestler's Predicted Win Percentage Against 8-6 Opponent	7-7 Wrestler's Actual Win Percentage Against 8-6 Opponent
48.7	79.6

So the 7–7 wrestler, based on past outcomes, was expected to win just less than half the time. This makes sense; their records in this tournament indicate that the 8–6 wrestler is slightly better. But in actuality, the wrestler on the bubble won *almost eight out of ten* matches

against his 8–6 opponent. Wrestlers on the bubble also do astonishingly well against 9–5 opponents:

7-7 Wrestler's Predicted Win Percentage Against 9-5 Opponent	7-7 Wrestler's Actual Win Percentage Against 9-5 Opponent
47.2	73.4

As suspicious as this looks, a high winning percentage alone isn't enough to prove that a match is rigged. Since so much depends on a wrestler's eighth win, he should be expected to fight harder in a crucial bout. But perhaps there are further clues in the data that prove collusion.

It's worth thinking about the incentive a wrestler might have to throw a match. Maybe he accepts a bribe (which would obviously not be recorded in the data). Or perhaps some other arrangement is made between the two wrestlers. Keep in mind that the pool of elite sumo wrestlers is extraordinarily tight-knit. Each of the sixty-six elite wrestlers fights fifteen of the others in a tournament every two months. Furthermore, each wrestler belongs to a stable that is typically managed by a former sumo champion, so even the rival stables have close ties. (Wrestlers from the same stable do not wrestle one another.)

Now let's look at the win-loss percentage between the 7–7 wrestlers and the 8–6 wrestlers the *next* time they meet, when neither one is on the bubble. In this case, there is no great pressure on the individual match. So you might expect the wrestlers who won their 7–7 matches in the previous tournament to do about as well as they had in earlier matches against these same opponents—that is, winning roughly 50 percent of the time. You certainly wouldn't expect them to uphold their 80 percent clip.

As it turns out, the data show that the 7–7 wrestlers win only 40 percent of the rematches. Eighty percent in one match and 40 percent in the next? How do you make sense of that?

The most logical explanation is that the wrestlers made a quid pro quo agreement: you let me win today, when I really need the victory, and I'll let you win the next time. (Such an arrangement wouldn't preclude a cash bribe.) It's especially interesting to note that by the two wrestlers' *second* subsequent meeting, the win percentages revert to the expected level of about 50 percent, suggesting that the collusion spans only two matches.

And it isn't only the individual wrestlers whose records are suspect. The collective records of the various sumo stables are similarly aberrational. When one stable's wrestlers fare well on the bubble against wrestlers from a second stable, they tend to do especially *poorly* when the second stable's wrestlers are on the bubble. This indicates that some match rigging may be choreographed at the highest level of the sport—much like the Olympic skating judges' vote swapping.

No formal disciplinary action has ever been taken against a Japanese sumo wrestler for match rigging. Officials from the Japanese Sumo Association typically dismiss any such charges as fabrications by disgruntled former wrestlers. In fact, the mere utterance of the words "sumo" and "rigged" in the same sentence can cause a national furor. People tend to get defensive when the integrity of their national sport is impugned.

Still, allegations of match rigging do occasionally find their way into the Japanese media. These occasional media storms offer one more chance to measure possible corruption in sumo. Media scrutiny, after all, creates a powerful incentive: if two sumo wrestlers or their stables *have* been rigging matches, they might be leery to continue when a swarm of journalists and TV cameras descend upon them.

So what happens in such cases? The data show that in the sumo tournaments held immediately after allegations of match rigging, 7–7 wrestlers win only 50 percent of their final-day matches against 8–6 opponents instead of the typical 80 percent. No matter how the data are sliced, they inevitably suggest one thing: it is hard to argue that sumo wrestling isn't rigged.

Several years ago, two former sumo wrestlers came forward with extensive allegations of match rigging—and more. Aside from the crooked matches, they said, sumo was rife with drug use and sexcapades, bribes and tax evasion, and close ties to the *yakuza*, the Japanese mafia. The two men began to receive threatening phone calls; one of them told friends he was afraid he would be killed by the *yakuza*. Still, they went forward with plans to hold a press conference at the Foreign Correspondents' Club in Tokyo. But shortly beforehand, the two men died—hours apart, in the same hospital, of a similar respiratory ailment. The police declared there had been no foul play but did not conduct an investigation. "It seems very strange for these two people to die on the same day at the same hospital," said Mitsuru Miyake, the editor of a sumo magazine. "But no one has seen them poisoned, so you can't prove the skepticism."

Whether or not their deaths were intentional, these two men had done what no other sumo insider had previously done: named names. Of the 281 wrestlers covered in the data cited above, they identified 29 crooked wrestlers and 11 who were said to be incorruptible.

What happens when the whistle-blowers' corroborating evidence is factored into the analysis of the match data? In matches between two supposedly corrupt wrestlers, the wrestler who was on the bubble won about 80 percent of the time. In bubble matches against a supposedly clean opponent, meanwhile, the bubble wrestler was no more likely to win than his record would predict. Furthermore, when a supposedly corrupt wrestler faced an opponent whom the whistle-

blowers did not name as either corrupt or clean, the results were nearly as skewed as when two corrupt wrestlers met—suggesting that most wrestlers who *weren't* specifically named were also corrupt.

So if sumo wrestlers, schoolteachers, and day-care parents all cheat, are we to assume that mankind is innately and universally corrupt? And if so, how corrupt?

The answer may lie in . . . bagels. Consider the true story of a man named Paul Feldman.

Once upon a time, Feldman dreamed big dreams. Trained as an agricultural economist, he wanted to tackle world hunger. Instead, he took a job in Washington, analyzing weapons expenditures for the U.S. Navy. This was in 1962. For the next twenty-odd years, he did more of the same. He held senior-level jobs and earned good money, but he wasn't fully engaged in his work. At the office Christmas party, colleagues would introduce him to their wives not as "the head of the public research group" (which he was) but as "the guy who brings in the bagels."

The bagels had begun as a casual gesture: a boss treating his employees whenever they won a research contract. Then he made it a habit. Every Friday, he would bring in some bagels, a serrated knife, and cream cheese. When employees from neighboring floors heard about the bagels, they wanted some too. Eventually he was bringing in fifteen dozen bagels a week. In order to recoup his costs, he set out a cash basket and a sign with the suggested price. His collection rate was about 95 percent; he attributed the underpayment to oversight, not fraud.

In 1984, when his research institute fell under new management, Feldman took a look at his career and grimaced. He decided to quit his job and sell bagels. His economist friends thought he had lost his

mind, but his wife supported him. The last of their three children was finishing college, and they had retired their mortgage.

Driving around the office parks that encircle Washington, he solicited customers with a simple pitch: early in the morning, he would deliver some bagels and a cash basket to a company's snack room; he would return before lunch to pick up the money and the leftovers. It was an honor-system commerce scheme, and it worked. Within a few years, Feldman was delivering 8,400 bagels a week to 140 companies and earning as much as he had ever made as a research analyst. He had thrown off the shackles of cubicle life and made himself happy.

He had also—quite without meaning to—designed a beautiful economic experiment. From the beginning, Feldman kept rigorous data on his business. So by measuring the money collected against the bagels taken, he found it possible to tell, down to the penny, just how honest his customers were. Did they steal from him? If so, what were the characteristics of a company that stole versus a company that did not? Under what circumstances did people tend to steal more, or less?

As it happens, Feldman's accidental study provides a window onto a form of cheating that has long stymied academics: white-collar crime. (Yes, shorting the bagel man is white-collar crime, writ however small.) It might seem ludicrous to address as large and intractable a problem as white-collar crime through the life of a bagel man. But often a small and simple question can help chisel away at the biggest problems.

Despite all the attention paid to rogue companies like Enron, academics know very little about the practicalities of white-collar crime. The reason? There are no good data. A key fact of white-collar crime is that we hear about only the very slim fraction of people who are *caught* cheating. Most embezzlers lead quiet and theoretically happy lives; employees who steal company property are rarely detected.

With street crime, meanwhile, that is not the case. A mugging or a

burglary or a murder is usually tallied whether or not the criminal is caught. A street crime has a victim, who typically reports the crime to the police, who generate data, which in turn generate thousands of academic papers by criminologists, sociologists, and economists. But white-collar crime presents no obvious victim. From whom, exactly, did the masters of Enron steal? And how can you measure something if you don't know to whom it happened, or with what frequency, or in what magnitude?

Paul Feldman's bagel business was different. It did present a victim. The victim was Paul Feldman.

When he started his business, he expected a 95 percent payment rate, based on the experience at his own office. But just as crime tends to be low on a street where a police car is parked, the 95 percent rate was artificially high: Feldman's presence had deterred theft. Not only that, but those bagel eaters knew the provider and had feelings (presumably good ones) about him. A broad swath of psychological and economic research has shown that people will pay different amounts for the same item depending on who is providing it. The economist Richard Thaler, in his 1985 "Beer on the Beach" study, showed that a thirsty sunbather would pay $2.65 for a beer delivered from a resort hotel but only $1.50 for the same beer if it came from a shabby grocery store.

In the real world, Feldman learned to settle for less than 95 percent. He came to consider a company "honest" if its payment rate was above 90 percent. He considered a rate between 80 and 90 percent "annoying but tolerable." If a company habitually paid below 80 percent, Feldman might post a hectoring note, like this one:

> The cost of bagels has gone up dramatically since the beginning of the year. Unfortunately, the number of bagels that disappear

without being paid for has also gone up. Don't let that continue. I don't imagine that you would teach your children to cheat, so why do it yourselves?

In the beginning, Feldman left behind an open basket for the cash, but too often the money vanished. Then he tried a coffee can with a money slot in its plastic lid, which also proved too tempting. In the end, he resorted to making small plywood boxes with a slot cut into the top. The wooden box has worked well. Each year he drops off about seven thousand boxes and loses, on average, just one to theft. This is an intriguing statistic: the same people who routinely steal more than 10 percent of his bagels almost never stoop to stealing his money box—a tribute to the nuanced social calculus of theft. From Feldman's perspective, an office worker who eats a bagel without paying is committing a crime; the office worker probably doesn't think so. This distinction probably has less to do with the admittedly small amount of money involved (Feldman's bagels cost one dollar each, cream cheese included) than with the context of the "crime." The same office worker who fails to pay for his bagel might also help himself to a long slurp of soda while filling a glass in a self-serve restaurant, but he is very unlikely to leave the restaurant without paying.

So what do the bagel data have to say? In recent years, there have been two noteworthy trends in the overall payment rate. The first was a long, slow decline that began in 1992. By the summer of 2001, the overall rate had slipped to about 87 percent. But immediately after September 11 of that year, the rate spiked a full 2 percent and hasn't slipped much since. (If a 2 percent gain in payment doesn't sound like much, think of it this way: the nonpayment rate fell from 13 to 11 percent, which amounts to a 15 percent decline in theft.) Because many of Feldman's customers are affiliated with national security,

there may have been a patriotic element to this 9/11 Effect. Or it may have represented a more general surge in empathy.

The data also show that smaller offices are more honest than big ones. An office with a few dozen employees generally outpays by 3 to 5 percent an office with a few hundred employees. This may seem counterintuitive. In a bigger office, a bigger crowd is bound to convene around the bagel table, providing more witnesses to make sure you drop your money in the box. But in the big-office/small-office comparison, bagel crime seems to mirror street crime. There is far less street crime per capita in rural areas than in cities, in large part because a rural criminal is more likely to be known (and therefore caught). Also, a smaller community tends to exert greater social incentives against crime, the main one being shame.

The bagel data also reflect how much personal mood seems to affect honesty. Weather, for instance, is a major factor. Unseasonably pleasant weather inspires people to pay at a higher rate. Unseasonably cold weather, meanwhile, makes people cheat prolifically; so do heavy rain and wind. Worst are the holidays. The week of Christmas produces a 2 percent drop in payment rates—again, a 15 percent increase in theft, an effect on the same magnitude, in reverse, as that of 9/11. Thanksgiving is nearly as bad; the week of Valentine's Day is also lousy, as is the week straddling April 15. There are, however, a few good holidays: the weeks that include the Fourth of July, Labor Day, and Columbus Day. The difference in the two sets of holidays? The low-cheating holidays represent little more than an extra day off from work. The high-cheating holidays are fraught with miscellaneous anxieties and the high expectations of loved ones.

Feldman has also reached some of his own conclusions about honesty, based more on his experience than the data. He has come to believe that morale is a big factor—that an office is more honest when the employees like their boss and their work. He also believes that

employees further up the corporate ladder cheat more than those down below. He got this idea after delivering for years to one company spread out over three floors—an executive floor on top and two lower floors with sales, service, and administrative employees. (Feldman wondered if perhaps the executives cheated out of an overdeveloped sense of entitlement. What he didn't consider is that perhaps cheating was how they got to *be* executives.)

If morality represents the way we would like the world to work and economics represents how it actually does work, then the story of Feldman's bagel business lies at the very intersection of morality and economics. Yes, a lot of people steal from him, but the vast majority, even though no one is watching over them, do not. This outcome may surprise some people—including Feldman's economist friends, who counseled him twenty years ago that his honor-system scheme would never work. But it would not have surprised Adam Smith. In fact, the theme of Smith's first book, *The Theory of Moral Sentiments*, was the innate honesty of mankind. "How selfish soever man may be supposed," Smith wrote, "there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it."

There is a tale, "The Ring of Gyges," that Feldman sometimes tells his economist friends. It comes from Plato's *Republic*. A student named Glaucon offered the story in response to a lesson by Socrates—who, like Adam Smith, argued that people are generally good even without enforcement. Glaucon, like Feldman's economist friends, disagreed. He told of a shepherd named Gyges who stumbled upon a secret cavern with a corpse inside that wore a ring. When Gyges put on the ring, he found that it made him invisible. With no one able to

monitor his behavior, Gyges proceeded to do woeful things—seduce the queen, murder the king, and so on. Glaucon's story posed a moral question: could any man resist the temptation of evil if he knew his acts could not be witnessed? Glaucon seemed to think the answer was no. But Paul Feldman sides with Socrates and Adam Smith—for he knows that the answer, at least 87 percent of the time, is yes.