The case against standardized testing: raising the scores, ruining the schools
By Alfie Kohn (2000)

Standardized testing has swelled and mutated, like a creature in one of those old horror movies, to the point that it now threatens to swallow our schools whole. Of course, on the late, late show no one ever insists that the monster is really doing us a favor by making its victims more "accountable." In real life, plenty of people need to be convinced that these tests do not provide an objective measure of learning or a useful inducement to improve teaching, that they are not only unnecessary but highly dangerous. This book was written to challenge those who defend the tests.

Other readers are already well aware of what is being sacrificed in the drive to raise scores, but they may find it helpful to have a few facts or research results at their fingertips, a quotable phrase or a set of answers to commonly asked questions. This book was written to assist those who oppose the tests.

Still others want for neither reasons nor rhetoric; what they lack is the requisite sense of urgency or the belief that they can make a difference. This book was written to energize and encourage those who have resigned themselves to the tests.

The more we learn about standardized testing, particularly in its high-stakes incarnation, the more likely we are to be appalled. And the more we are appalled, the more inclined we will be to do what is necessary to protect our children from this monster in the schools.

MEASURING WHAT MATTERS LEAST

Is it my imagination, or are we spending an awful lot of time giving kids standardized tests?

It's not your imagination. While previous generations of American students have had to sit through tests, never have the tests been given so frequently **and** never have they played such a prominent role in schooling. Exams used to be administered mostly to decide where to place kids or what kind of help they needed; only recently have scores been published in the newspaper and used as the primary criteria for judging children, teachers, and schools—indeed, as the basis for flunking students or denying them a diploma, deciding where money should be spent, and so on. Tests have lately become a mechanism by which public officials can impose their will on schools, and they are doing so with a vengeance.

This situation is also unusual from an international perspective. "Few countries today give these formal examinations to students before the age of sixteen or so," two scholars report. In the U.S., we subject children as young *as six* to standardized exams, despite the fact that almost all experts in early childhood education condemn this practice. And it isn't easy to find other countries that give multiple-choice tests to students of any age.

In short, our children are tested to an extent that is unprecedented in our history and unparalleled anywhere else in the world. Rather than seeing this as odd, or something that needs to be defended, many of us have come to take it for granted. The result is that most of today's discourse about education has been reduced to a crude series of monosyllables: "Test scores are too low. Make them go up."

So what accounts for this?

Well, different people have different motivations. For some, a demand for tests seems to reflect a deliberate strategy for promoting traditional, "back-to-basics" instruction. (Whether or not that's the intent, it's often the consequence of an emphasis on standardized test scores.) Other people, meanwhile, are determined to cast public schools in the worst possible light as a way of paving the way for the privatization of education. After all, if your goal was to serve up our schools to the marketplace, where the point of reference is what maximizes profit rather than what benefits children, it would be perfectly logical for you to administer a test that many students would fail in order to create the impression that public schools were worthless. Not everyone has ulterior motives for testing, of course. Some people just insist that schools have to be held "accountable," and they don't know any other way to achieve that goal. Even here, though, it's worth inquiring into the sudden, fierce demands for accountability. The famous *Nation at Risk* report released by the Reagan Administration in 1983 was part of a concerted campaign—based on exaggerated and often downright misleading evidence—to stir up widespread concerns about our schools and, consequently, demands for more testing.

There's another built-in constituency: the corporations that manufacture and score the exams, thereby reaping enormous profits (on revenues estimated at nearly a quarter of a *billion* dollars in 1999, and continuing to grow rapidly). More often than not, these companies then turn around and sell teaching materials designed to raise scores on their own tests. The worst tests are often the most appealing to school systems: It is fast, easy, and therefore relatively inexpensive to administer a multiple-choice exam that arrives from somewhere else and is then sent back to be graded by a machine at lightning speed. There is little incentive to replace these tests with more meaningful forms of assessment that require human beings to evaluate the quality of students' accomplishments. "Efficient tests tend to drive out less efficient tests, leaving many important abilities untested—and untaught."

Testing allows politicians to show they're concerned about school achievement and serious about getting tough with students and teachers. Test scores offer a quick-and-easy—although, as we'll see, by no means accurate—way to chart progress. Demanding high scores fits nicely with the use of political slogans like "tougher standards" or "accountability" or "raising the bar."

If the public often seems interested in test results, it may be partly because of our cultural penchant for attaching numbers to things. Any aspect of learning (or life) that appears in numerical form seems reassuringly scientific; if the numbers are getting larger over time, we must be making progress. Concepts such as intrinsic motivation and intellectual exploration are difficult for some minds to grasp, whereas test scores, like sales figures or votes, can be calculated and tracked and used to define success and failure. Broadly speaking, it is easier to measure efficiency than effectiveness, easier to rate how well we're doing something than to ask whether what we're doing makes sense. Not everyone realizes that the process of coming to

understand ideas in a classroom is not always linear or quantifiable—or, in fact, that "measurable outcomes may be the least significant results of learning."

But don't we need an objective measure of achievement?

This question is much more complicated than it may appear. Is objectivity really a desirable—or a realistic—goal? Presumably, an "objective" assessment is one that's not dependent on subjective factors such as the beliefs and values of different individuals; everyone would have to agree that something was good *or* bad. But disagreement is a fact *of life,* and it isn't necessarily something to be transcended. You and I will inevitably differ in our judgments about politics and ethics, about the quality of the movies we see and the meals we eat. It is odd and troubling that in educating our children "we expect a different standard of assessment than is normal in the rest of our lives."

Too much standardization suggests an *effort* to pretend that evaluations aren't ultimately judgments, that subjectivity can be overcome. This is a dangerous illusion. Testing specialists always seem to be chasing the holy grail of "inrerrater reliability," but there's no reason to expect that people will always see eye-to-eye about the value of what students have done. If they do, that suggests either that they have obediently set aside their own judgments in order to rigidly apply someone else's criteria, or that the assessment in question is fairly superficial. For example, it's easier to get agreement on whether a semicolon has been used correctly than on whether an essay represents clear thinking. The quest for objectivity may lead us to measure students on the basis of criteria that are a lot less important.

For the sake of the argument, though, let's assume that objective assessments are both possible and desirable. The critical point is that *standardized tests do not provide such objectivity.* It's easy to assume otherwise when a precise numerical score has been assigned to a student or school. But the testing process is nothing at all like, say, measuring the size or weight of an object. The results may sound scientific, but they emerge from the interaction of two sets of human beings: the invisible adults who make up the questions and the rows of kids, scrunched into desks, frantically writing (or filling in bubbles).

First, we need to know about the content of the test. Are we measuring something important? One can refer to it as "objective" in the sense that it's scored by machines, but people wrote the questions (which may be biased or murky or stupid) and people decided to include them on the exam. Reasonable doubts often can be raised about which answers ought to be accepted, even at the elementary school level, where you might expect the questions to be more straightforward. Thus, to read narrative accounts of students who think through a given question and arrive at a plausible answer—only to learn that the answer has been coded as incorrect is to understand the limits of these putatively objective assessment instruments.

The significance of the scores becomes even more dubious once we focus on the experience of students. For example, test anxiety has grown into a subfield of educational psychology, and its prevalence means that the tests producing this reaction are not giving us a good picture of what many students really know and can do. The more a test is made to "count"—in terms of being the basis for promoting or retaining students, for funding or closing down schools—the more that anxiety is likely to rise and *the less valid the scores become.*

Then there are the students who take the tests but don't take them seriously. A friend of mine remembers neatly filling in chose ovals with his pencil in such *a way* that they made a picture of a Christmas tree. (He was assigned to a low-level class as a result, since his score on a single test

was all the evidence anyone needed of his capabilities.) Even those test-takers who are not quite so creative may just guess wildly, fill in ovals randomly, or otherwise blow off the whole exercise, understandably regarding it as a waste of time. In short, it may be that a good proportion of students either couldn't care less about the tests, on the one hand, or care so much that they choke, on the other. Either way, the scores that result aren't very meaningful. Anyone who can relate to these descriptions of what goes through the minds of real students on test day ought to think twice before celebrating a high score, complaining about a low one, or using standardized tests to judge schools.

Even if they're not "objective," though, wouldn't you agree that we need some way to tell which students are ready for the world of work? It's just not realistic to think we can eliminate testing.

The more you're concerned about what's "realistic," the more critical you should be of standardized tests. How many jobs demand that employees come up with the right answer on the spot, from memory, while the clock is ticking? (I can think of one or two, but they're the exceptions that prove the rule.) How often are we forbidden to ask coworkers for help, or to depend on a larger organization for support—even in a society that worships self-sufficiency? And when someone is going to judge the quality of your work, whether you are a sculptor, a lifeguard, a financial analyst, a professor, a housekeeper, a refrigerator repairman, a reporter, or a therapist, how common is it for you to be given a secret pencil-and-paper exam? Isn't it far more likely that the evaluator will look at examples of what you've already done, or perhaps watch you perform your normal tasks? To be consistent, those educational critics who indignantly insist that schools should be doing more to prepare students for the real world ought to be demanding an end to these artificial exercises called standardized tests.

Including the college admission tests, the SATs and ACTs?

Ideally, yes. These tests are not very effective as predictors of future academic performance, even in the freshman year of college, much less as predictors of professional success. They're not good indicators of thinking or aptitude; the verbal section is basically just a vocabulary test. (The "A" in SAT used to stand for Aptitude until the Educational Testing Service gave up this pretense. Now "SAT" doesn't stand for anything—in more ways than one.) They're not necessary for deciding who should *be* admitted to college. No such exams are used in Canada, for example, and several hundred U.S. colleges and universities no longer require applicants to take them.

But these tests are not our primary concern here. It's far more worrisome that even students who don't plan to continue their schooling after high school, and even students who are much too young to be thinking about college are subjected to a barrage of standardized tests that don't provide much useful information.

The results of these tests must tell us something.

The main thing they tell us is how big the students' houses are. Research has repeatedly found that the amount of poverty in the communities where schools are located, along with other variables having nothing to do with what happens in classrooms, accounts for the great majority of the difference in test scores from one area to the next. To that extent, tests are simply not a valid measure of school effectiveness. (Indeed, one educator suggested that we

could save everyone a lot of time and money by eliminating standardized tests and just asking a single question: "How much money does your mom make? ... OK, you're on the bottom.") Only someone ignorant or dishonest would present a ranking of schools' test results as though it told us about the quality of teaching that went on in chose schools when, in fact, it primarily tells us about socioeconomic status and available resources. Of course, knowing what really determines the scores makes it impossible to defend the practice of using them as the basis for high-stakes decisions.

But socioeconomic status isn't everything. Within a given school, or group of students of the same status, aren't there going to be variations in the scores?

Sure. And among people who smoke three packs of cigarettes a day, there are going to be variations in lung cancer rates. But that doesn't change the fact that smoking is the factor most powerfully associated with lung cancer.

Still, let's put wealth aside and just focus on the content of the tests themselves. The fact is that they usually don't assess the skills and dispositions that matter most. They tend to be contrived exercises that measure how much students have managed to cram into short-term memory. Even the exceptions—questions that test the ability to reason—generally fail to offer students the opportunity "to carry out extended analyses, to solve open-ended problems, or to display command of complex relationships, although these abilities are at the heart of higher-order competence," as Lauren Resnick, one of our leading cognitive scientists, put it.

Part of the problem rests with an obvious truth whose implications we may not have considered: These tests care only about whether the student got the right answer. To point this out is not to claim that there is no such thing as a right answer; it is to observe that right answers don't necessarily signal understanding, and wrong answers don't necessarily signal the absence of understanding. Most standardized tests ignore the process by which students arrive at an answer, so a miss is as good as a mile and a minor calculation error is interchangeable with a major failure of reasoning.

The focus on right answers also means that most, if not all, of the items on the test were chosen precisely because they have unambiguously correct solutions, with definite criteria for determining what those solutions are and a clear technique for getting there. The only thing wrong with these questions is that they bear no resemblance to most problems that occupy people in the real world.

You're making some sweeping statements here. What subjects are you talking about? Reading? Math?

You pick. What generally passes for a test of reading comprehension is a series of separate questions about short passages on random topics. These questions "rarely examine how students interrelate parts of the text and do not require justifications that support the interpretations"; indeed, the whole point is the "quick finding of answers rather than reflective interpretation."

In mathematics, the story is much the same. An analysis of the most widely used standardized math rests found that only *3* percent of the questions required "high level conceptual knowledge" and only 5 percent tested "high level thinking skills such as problem solving and reasoning." Typically the rests aim to make sure that students have memorized *a*

series of procedures, not that they understand what they are doing. They also end up measuring knowledge of arbitrary conventions (such as the accepted way of writing a ratio or the fact that "<" means "less than") more than a capacity for logical thinking. Even those parts of math rests that have names like "Concepts and Applications" are "still given in multiple-choice format, *are* computational in nature, and test for knowledge *of* basic skills through the use of traditional algorithms."

The parts of standardized exams that deal with science or social studies, meanwhile, typically amount to nothing more than a test of obscure facts and definitions. They aren't designed to tell who has learned to think like a scientist or an historian; they're designed to tell who can recite the four stages of mitosis or the four freedoms mentioned by Franklin Roosevelt. As the president of the National Academy of Sciences has remarked, questions that focus on "excruciatingly boring material" not only fail to judge students' capacity to reason but wind up driving away potential future scientists."

Are you saying the tests are too hard?

Sometimes they are. Plenty of successful adults would fail the high school exit exams used in many states—and might not even do all that well on some of the tests given to fourth graders. But the real problem isn't the difficulty level, per se. It's the fact that these tests are geared to a different, less sophisticated kind *of* knowledge. It's not just that the tests are often ridiculously hard; it's that they're simply ridiculous. They don't capture what most of us, upon reflection, would say it means to be a well-educated person.

Two math educators offer a good example from the Massachusetts test for high school students. The question reads as follows:

n 1 2 3 4 5 6

t_n 3 5

The first two terms of a sequence, t, and t_2, are shown above as 3 and 5. Using the rule:

$t_n = t_{n-2} + t_{n-2}$, where n is greater than or equal to 3, complete the table.

This is actually just asking the test taker to add 3 and 5 to get 8, then add 5 and 8 to get 13, then add 8 to 13 to get 21, and so on.

The problem simply requires the ability to follow a rule; there is no mathematics in it at all. And many tenth-grade students will get it wrong, not because they lack the mathematical thinking necessary to fill in the table, but simply because they haven't had experience with the notation. Next year, however, teachers will prep students on how to use formulas like $t_n = t_{n-1+} t_{n-2,}$ more students will get it right, and stare education officials will tell us that we are increasing mathematical literacy.

Even if the tests are imperfect, don't top students still do the best?

That depends on what you mean by "top students." If you mean those who are most interested in learning and most likely to think deeply, then the answer may surprise you. Although these findings haven't been widely publicized, studies of students of different ages have found *a statistical association between high scores on standardized tests and relatively shallow thinking.* One of these studies classified elementary school students as "actively" engaged in learning if *they* went back over things they didn't understand, asked questions of themselves as they read, and tried to connect what they were doing to what they had already learned; and as

"superficially" engaged if they just copied down answers, guessed a lot, and skipped the hard parts. It turned out that the superficial style was positively correlated with high scores on the Comprehensive Test of Basic Skills (CTBS) and Metropolitan Achievement Test (MAT). Similar findings have emerged from studies of middle school and high school students.

These are only statistical relationships, you understand—significant correlations, but not absolute correspondences. There are plenty of students who think deeply *and* score well on tests. There are also plenty of students who do neither. But as a rule, good standardized test results are more likely to go hand in hand with a shallow approach to learning than with deep understanding. By virtue of their design (more about which later), "most tests *punish* the thinking test-taker"—to the point that some teachers advise their students, in effect, to dumb themselves down so they can do better on the tests.

Perhaps this is why, as Piaget pointed out years ago, "Anyone can confirm how little the grading that results from examinations corresponds to the final useful work of people in life." But never mind their inability to predict what students will be able to do later; they don't even capture what students can do today. In fact, we might say that such tests fail in two directions at once. On the one hand, they overestimate what some students know: Those who score well often understand very little of the subject in question. Students may be able to find a synonym or antonym for a word without being able to use it properly in a sentence. They may have memorized the steps of comparing the areas of two figures without really understanding geometric principles at all. On the other hand, standardized tests underestimate what others can do because, as any teacher can tell you, very talented students often get low scores. For example, there are "countless cases of magnificent student writers whose work was labeled as not proficient' because it did not follow the step-by-step sequence of what the test scorers (many of whom are not educators, by the way) think good expository writing should look like."

Consider a fifth grade boy who, researchers found, could flawlessly march through the steps of subtracting 2 5/6 from 3 1/3, ending up quite correctly with 3/6 and then reducing that to 1/2. Unfortunately, successful performance of this final reduction does not imply understanding that the two fractions are equivalent. In fact, this student remarked in an interview that 1/2 was larger than 3/6 because "the denominator is smaller so the pieces are larger." Meanwhile, one of his classmates, whose answer had been marked wrong because it hadn't been expressed in the correct terms, clearly had a better grasp of the underlying concepts. Intrigued, these researchers proceeded to interview a number of fifth graders about another topic (division) and discovered that 41 percent had memorized the process without really understanding the idea, while 11 percent understood the concept but made minor errors that resulted in getting the wrong answers. A standardized test therefore would have misclassified more than half of these students.

THE WORST TESTS

Surely, though, all *standardized tests aren't this bad.*

No, some are even worse. The most damaging testing programs are characterized by certain readily identifiable features, beginning with the use of exams that are mostly **multiple choice**.

"I don't think there's any way to build a multiple-choice question that allows students to show what they can do with what they know," says Roger Farr, professor of education at Indiana University—a statement all the more remarkable given that Farr personally helped to write a number of standardized tests. The reasons should be obvious. Students are unable to *generate* a response; all they can do is recognize one by picking it out of four or five answers provided by someone else. They can't even *explain* their reasons for choosing the answer they did. Obviously some sort of remembering, calculating, or thinking has to be done to figure out which answer is "most appropriate," but other sorts of mental operations (such as organizing information or constructing an argument) are pretty much excluded by the format. No matter how clever or tricky the questions are, a multiple-choice test simply "does not measure the same cognitive skills as are measured by similar problems in free-response form," as one expert explained in a now-classic article. The difference between the two formats (which is to say, the limits of multiple-choice questions) really shows up when the idea is to measure "complex cognitive problem-solving skills."

Well, I'm relieved. My state's exam has a lot of multiple-choice questions on it, but at least it has some open-response items, too.

Unfortunately, even essay questions often leave a lot to be desired. They may require students to analyze a dull chunk of text, cough up obscure facts, or produce cogent opinions on command about some bland topic—hardly an authentic assessment of meaningful learning. What's more, these questions are often scored on the basis of imitating a contrived model (such as a cookie-cutter five-paragraph essay) rather than tapping real communication or thinking skills. Preparing kids to turn out high-scoring essays can *inhibit* the quality of their writing.

The way these exams are graded raises even more concerns. For example, the essays written by students in many states are not evaluated by educators; *they* are shipped off to a company in North Carolina where low-paid temp workers spend no more than two or three minutes reading each one. "There were times I'd be reading a paper every ten seconds," one former scorer told a reporter. Sometimes he "would only briefly scan papers before issuing a grade, searching for clues such as a descriptive passage within a narrative to determine what grade to give. `You could skim them very quickly ... I know this sounds very bizarre, but you could put a number on these things without actually reading the paper,'" said this scorer, who, like his coworkers, was offered a "two hundred dollar bonus that kicked in after eight thousand papers." In short, we can't assume that an essay test is a valid measure of important things. But we can be reasonably certain that a multiple-choice test *isn't.*

All right, what else is relevant besides the format of the questions?

First, beware of tests that are timed. If students must complete an examination within a specified period, this means that a premium is placed on speed as opposed to thoughtfulness or even thoroughness. If one small part of the test were timed, this would indicate that the ability to do things quickly and under pressure was one of several valued attributes. But if the entire exam must be taken under the gun, the logical inference is that this ability is prized above others.

Second, you should be worried if tests are given frequently. It is neither necessary (in terms of collecting information) nor desirable (in terms of improving the quality of instruction) to test students year after year after year. This practice is generally connected to grade-by-grade performance standards, and they, in turn, reflect the assumption that all students must learn at the same pace. As a descriptive premise, this is out of step with developmental reality; as a prescriptive formula, it ensures that those who require more time to learn will be branded as failures. The uniformity implied in grade-level standards and testing emphasizes the speed (measured in months or years) at which students must master a set curriculum—an interesting echo of the disproportionate emphasis on speed (measured in minutes or hours) reflected in the use of timed exams.

Third, be prepared to protest if tests are given to **young children**. Students below fourth grade simply should not be subjected to standardized examinations—first, because it is difficult, if not impossible, to devise such an assessment in which they can communicate the depth of their understanding; and second, because skills develop rapidly and differentially in young children, which means that expecting all second graders to have acquired the same skills or knowledge creates unrealistic expectations and leads to one-size-fits-all (which is to say, poor) teach-ing. In fact, "what test-makers are measuring for some children" is not their cognitive capacities so much as their "ability to sit in the same place for a certain amount of time."

Finally, look out for tests that are **"norm-referenced."**

I've heard that term a lot, but I've never understood exactly what it means.

Robert Glaser coined the term "norm-referenced test" (NRT) many years ago to refer to tests that "provide little or no information about …. what the individual can do. They tell that one student is more or less proficient than another, but do not tell how proficient either of them is with respect to the subject matter tasks involved." The most common norm-referenced tests are the Iowa and Comprehensive Tests of Basic Skills (ITBS and CTBS), and the Stanford, Metropolitan, and California Achievement Tests (SAT, MAC, and CAT). In contrast to a test that's "criterion-referenced," which means it compares each individual to a set standard, one that's norm-referenced compares each individual to everyone else, and the result is usually (but not always) reported as a percentile.

Think for a moment about the implications of this. No matter how many students take an NRT, no matter how well or poorly they were taught, no matter how difficult the questions are, the pattern of results is guaranteed to be the same: Exactly 10 percent of those who take the test will score in the top 10 percent, and half will always fall below the median. That's not because our schools are failing; that's because of what the word *median* means. A good score on an NRT means "better than other people," but we don't even know how much better. It could be that everyone's actual scores are all pretty similar, in which case the distinctions between them are meaningless, rather like saying I'm the tallest person on my block even though I'm only half an inch taller than the shortest person on my block.

More important, even if the top 10 percent did a *lot* better than the bottom 10 percent, that still doesn't tell us anything at all about how well they did in absolute terms, such as how many

questions they got right. Maybe everyone did reasonably well; maybe everyone blew it. We don't know. Norm-referenced tests cannot tell us—indeed, were never designed to tell us—how much of a body of knowledge a student learned or a school taught. To try to use them for those purposes is, in the words of W. James Popham, a leading authority, "like measuring temperature with a tablespoon." Yet NRTs *are* used for exactly those purposes all across the United States, often by people who should know better.

Norm-referenced tests are not about assessing excellence; they are about sorting students (or schools) into winners and losers. The animating spirit is not "How well are *they* learning?" but "Who's beating whom?" The latter question doesn't provide useful information because the only thing that really counts is how many questions on a test were answered correctly (assuming they measured important knowledge). By the same token, the news that your state moved up this year from thirty-seventh in the country to eighteenth doesn't tell us whether its schools are really improving: for all you know, the schools in your state are in worse shape than they were last *year*, but those in other states slid even further.

Even that isn't the whole story. When specialists sit down to construct an NRT they're not interested in making sure the questions cover what is most important for students to know. Rather, their goal is to include questions that some test-takers—not all of them, and not none of them—will get right. They don't *want* everyone to do well on the test. The ultimate objective, remember, is not to evaluate how well the students were taught, but to separate them, to get a range of scores. If a certain question is included in a trial test and almost everyone gets it right—or, for that matter, if almost no one gets it right—that question will likely be tossed out. Whether it is reasonable for kids to get it right is irrelevant.

Even if these tests aren't as informative as we've been led to believe, what's the harm of seeing how kids stack up against one another?

Given that scores from NRFs are widely regarded as if *they* contained meaningful information about how our children (and their schools) are doing, they are not only dumb but dangerous. And the harm ramifies through the whole system in a variety of ways. First, these tests contribute to the already pathological competitiveness of our culture, where we come to regard others as obstacles to our own success—with all the suspicion, envy, self-doubt, and hostility that rivalry entails. The process of assigning children to percentiles helps to ensure that schooling is more about triumphing over everyone else than it is about learning.

Second, because every distribution of scores contains a bottom, it will always appear that some kids are doing terribly.

That, in turn, reinforces a sense that the schools are failing. Worse, it contributes to the insidious assumption that some children just can't learn—especially if the same kids always seem to show up below the median. (This conclusion, based on a misunderstanding of statistics, is then defended as "just being realistic.") Parents and teachers may come to believe this falsehood, and so too may the kids themselves. They might figure: No matter how much I improve, everyone else will probably get better too, and I'm always going to be at the bottom. So why bother trying? Conversely, a very successful student, trained to believe that rankings are what matter, may be confident of remaining at the top and therefore have no reason to do as well as possible. (Excellence and victory, after all, are two completely different goals.) For both groups of students, it is difficult to imagine a more powerful demotivator than norm-referenced testing.

There's more: The questions that "too many" students will answer correctly probably are those that deal with the content teachers have *been* emphasizing in class because they judged it to be important. So NRTs are likely to include a lot of trivial stuff that *isn't* emphasized in school because that material is useful for distinguishing one student from another. Therefore, teachers and administrators who are determined to outsmart the test—or who are under pressure to bring up their school's rank—may try to adjust the curriculum in order to bolster their students' scores. But if the tests emphasize relatively unimportant knowledge that's designed for sorting, then "teaching to the test" isn't going to improve the quality of education. It may have exactly the opposite effect.

These basic facts should be understandable to almost everyone, yet the mind boggles at the reality that our children continue to be subjected to tests like the ITBS and the current version of the Stanford Achievement Test (the SAT-9), which are both destructive and ridiculously ill-suited to the purposes for which they are used.

So I can relax if my state's test isn't norm-referenced?

All else being equal, a test is certainly less damaging if it's not set up as a zero-sum game. Nevertheless, these tests may be treated as though they *were* norm-referenced. That can happen *if* parents or students aren't helped to understand that a score *of* 80 percent refers to the proportion of questions answered correctly, leaving them to assume that it refers to a score better than 80 percent of the other test-takers. Worse yet, criterion-referenced tests *may* be turned *into* the norm-referenced kind if newspapers publish charts showing how every school or district ranks on the same exam, thereby calling attention to what is least significant. (One expert on testing suggests that if newspapers insist on publishing such a chart, they should at least place it where it belongs, in the sports section.)

Finally, even if your state officials know better than to subject kids to NRTs, your local officials may not. Plenty of school districts are making students take norm-referenced (mostly multiple-choice) tests on top of the ones mandated by the state.

It's starting to sound as though you don't like any standardized tests.

Again, not all tests are equally bad. The least useful and most damaging testing program would be one that uses (1) a norm-referenced exam in which students must answer (2) multiple-choice questions in a (3) fixed period of time—and must do so (4) repeatedly, beginning when they are (5) in the primary grades. But remember: Even testing programs that avoid some or all of these pitfalls are likely to be problematic to the extent *they* measure mere memorization or even test-taking skills. In any case, *all* standardized tests tend to ignore the most important characteristics of a good learner, to say nothing of a good person. Here's a list offered by educator Bill Ayers, although you might just as well make up your own:

> Standardized tests can't measure initiative, creativity, imagination, conceptual thinking, curiosity, effort, irony, judgment, commitment, nuance, good will, ethical reflection, or a host of other valuable dispositions and attributes. What they can measure and count are isolated skills, specific facts and functions, the least interesting and least significant aspects of learning.

Beyond their ineffectiveness as assessments, note that the act of administering (and emphasizing the results of*)* standardized tests can communicate some pointed lessons about the

nature of learning. Because there is a premium placed on remembering facts, children may come to think that this is what really matters—and they may even come to develop a "quiz show" view of intelligence that confuses being smart with knowing a lot of stuff. Because the tests are timed, students may be encouraged to see intelligence as a function of how quickly people can do things. Because the tests often rely on a multiple-choice format, students may infer "that a right or wrong answer is available for all questions and problems" in life and that "someone else already knows the answer to [all these questions], so original interpretations are not expected; the task is to find or guess the right answer, rather than to engage in interpretive activity."

Two other features of standardized tests also may teach dubious lessons even as they detract from the tests' usefulness. First, they're given to individuals, not to groups, and helping one another is regarded as a serious offense. Not only is there no measure of the capacity to cooperate effectively, or even to assimilate other people's ideas into your own, but precisely the opposite message is communicated: Only what you can do alone is of any value. "We have been so convinced of the notion that intellect is an isolated, individual quality that we utterly lack the procedures or the psychometrics to study students' performances in group situations," as Dennie Wolf and her colleagues put it.

Second, the content of these tests is kept secret. Given their nature, this is hardly surprising, but look at it this way: What does it say about an approach to assessment that it can be done only by playing "Gotcha!"? Tests "that continually keep students in the dark are built upon procedures with roots in pre-modern traditions of legal proceedings and religious inquisitions." Apart from raising stress levels, the kind of evaluation where students aren't allowed to know in advance what they'll be asked to do suggests a heavy emphasis on memorization. It also has the practical effect of preventing teachers from reviewing the test with students after it's over and using it as a learning tool.

I'm sorry, but I just don't see how you could have a standardized test that didn't *have right answers, or wasn't secret or timed or whatever.*

You probably couldn't. That's my point. Many of the problems identified here are inherent to standardized testing. But here is a very different question: Could you devise a way of figur-ing out how well students are learning, or teachers are teaching, that didn't have these features? As we'll see, this question does have an answer. But it's critical that we frame the issue in these broader terms so that this becomes our point of departure. Only then are we free to look beyond—and avoid the problems created by—standardized tests.

BURNT AT THE HIGH STAKES

Do most people in the field of education recognize the problems you've described here?

There are no data on this, but my impression is that the people who work most closely with kids are the most likely to understand the limits of standardized tests. An awful lot of teachers—particularly those who are very talented—have what might be described as a dislike/hate relationship with testing. But support for testing seems to grow as you move away from the students, going from teacher to principal to central office administrator to school board member to state board member, state legislator, and governor. Those for whom

classroom visits are occasional photo opportunities are most likely to be big fans of testing and to offer self-congratulatory sound bites about the need for accountability.

But what happens when teachers or students explain that they'd rather pursue other kinds of learning, that they don't care about scores? Doesn't this lead the people in charge to rethink the value of the tests?

To the contrary, most of them have responded by saying, in effect, "Well, then, we'll *force you* to care about the scores!" This they have done in several ways: first, by making sure the tests are given frequently, raising their visibility among teachers and students; second, by publishing the scores and encouraging the public to see them as indicators of school quality--even hoping that bad results might serve as a kind of "public shaming" that will pressure educators to do anything necessary to crank up their scores.

Finally, officials have responded by using an assortment of bribes and threats to coerce everyone into concentrating on the test results. If the scores are high, the bribes may include bonuses for teachers and schools. Students, meanwhile, may receive food, tickets to theme parks or sporting events, exemptions from in-class final exams, and even substantial scholarships. The threats include loss of funding or accreditation for schools, while students may be held back a year or denied a high school diploma if they don't test well, regardless of their overall academic record. Collectively, these kinds of tactics are known as "high-stakes" testing.

Some of these methods do seem harsh, but doesn't it make sense to put some teeth" into the standards?

Not unless you think the way to improve education is by biting people. These policies are unwise for many reasons, beginning with the deficiencies of the tests themselves. Always remember that it is the results on those deeply flawed exams that determine who gets rewarded or punished. In fact, some states and cities are actually making rewards and punishments contingent on the results of *norm-referenced* tests (e.g., the ITBS in Chicago, the Stanford-9 in California), a policy that has rightly been described as educational malpractice.

Before we look at the real-world effects of high-stakes testing, it's worth considering that the approach is simply unfair. It holds people "accountable" for factors over which they have little control, which is as pointless as it is cruel. For example, low scores—in absolute and especially in relative terms—are to a large extent due to social and economic factors, as we've already seen. Those factors include the resources available to the school as well as the level of affluence of the community in which the school is located. But even to the extent that the scores do reflect school experience, that experience is hardly lim-ited to the current year. Thus, it seems difficult to justify holding a fourth-grade teacher accountable for her students' test scores when those scores reflect all that has happened to the children before they even arrived at her class.

Then there is the possibility for error, which becomes far more disturbing when high stakes are attached to test results. It seems as though every month or so one of the big test publishers, Harcourt Brace, CTB/McGraw Hill, or Riverside, makes some sort of mistake scoring exams. In one such episode, New York City officials ordered 8,600 students to attend remedial summer school on the basis of a scoring error on the CTBS. Still more unsettling is the fact that standardized tests have inaccuracies built into them. Even when they are scored correctly, and even when they meet conventional standards for reliability, many children will be misclassified

because of the limits of test accuracy. A Stanford University researcher calculated that a student whose hypothetical "real achievement" is at the fiftieth percentile will actually score within five percentage points of that level only about 30 percent of the time on the SAT-9 math exam and 42 percent of the time on the reading exam. Yet rewards and punishments hinge on such scores as though they were perfect measures of achievement.

But the basic idea of giving people an incentive to improve makes sense, doesn't it?

Not really. A detailed explanation of this point would take us too far afield, but suffice it to say that rewards and punishments can never succeed in producing more than temporary compliance, and even that result is achieved at a substantial cost.

People can sometimes be dissuaded from doing certain things if they are threatened with a punitive consequence, but this tends to create a climate of fear, which, in turn, generates anger and resentment. It also leads people to act more cautiously. As a rule, human beings are less likely to think creatively when they perceive themselves to be under threat. (Hence the wry humor of a sign posted in some offices and classrooms, which could be the motto of the contemporary "tougher standards" movement: TIIE BEATINGS WILL CONTINGE UNTIL MORALE IMPROVES.) When individuals are threatened with the deprivation of money, status, autonomy, or something else they value, any temporary effect in the desired direction—desired, that is, by the individual with the power to issue these threats—is usually more than offset by the demoralization that occurs.

The use of punishments and threats is sometimes justified on the grounds that, *however disagreeable,* it *succeeds* in "motivating" people. But this argument is based on the simplistic and ultimately faulty assumption that motivation consists of a single entity that people possess to a greater or lesser degree: Threaten someone with an aversive consequence unless she does x, and her motivation to do *it* will rise. Decades of psychological theory and research have challenged this view by demonstrating that there are different kinds of motivation. Moreover, it appears that *the kind matters more than the amount.* Psychologists typically distinguish between "intrinsic" and "extrinsic" *motivation,* depending upon *whether one sees* a task as valuable in its own right or merely a means to an end. It's obvious to most of us that these two forms of motivation are qualitatively different. It's also reasonably clear that intrinsic motivation is more desirable and more potent over the long haul. No amount *of extrinsic* motivation to do something *can* compensate for an absence of genuine enthusiasm. Adults who consistently do excellent work, and students whose learning is most impressive, are usually those who love what they do, not those who see what they do as a way to escape a punishment (such as losing out on a bonus or being forced to *repeat* a grade).

Furthermore, extrinsic motivation is not merely different or inferior; *it's corrosive.* That is, *it* rends to undermine intrinsic motivation. Under most real-life conditions, these two forms of motivation are likely to be reciprocally related. Someone acting to avoid a punishment is apt to lose interest in that which he was threatened into doing. Teaching and learning alike come to be seen as less appealing when someone has a gun to your head.

But what if no punishments are used? What if someone is just offered a reward for doing a good job?

That, too, is a form of extrinsic motivation. In fact, there's even more evidence about the destructive effects of rewards than there is about punishments. Scores of studies have demon-strated that *the more people are rewarded for doing something, the more they tend to lose interest in whatever they had to do to get the reward*. Thus, the intrinsic motivation that is so vital to quality—to say nothing of quality of life—often evaporates in the face of extrinsic incentives, be they carrots or sticks.

Rewards and punishments are sometimes described as though they were opposites—and as though they exhausted the available strategies for effecting change. Thus, discussions are framed in terms of which one is preferable. The truth of the matter is that the two are mirror images of one another, variations on a single theme. Both represent ways of doing things *to* people, as distinct from working *with* people. Indeed, one reason that extrinsic inducements are likely to be counterproductive is that they are widely, and usually correctly, construed as tactics of control. This is more overt in the case of punishment ("Do this or here's what will be done to you") but no less true in the case of rewards ("Do this and you'll get that"). The more desirable the incentive, the more that using it to get people to act in a particular way is likely to backfire, particularly when the goal is something deeper, more complex, or longer lasting than temporary compliance.

The more familiar one becomes with the psychological research, the sillier one realizes it is to use rewards or punishments as a way of "motivating" people to accomplish important goals.

But most of that research didn't deal with high-stakes testing right?

That's correct. Here my point has been only that the psycho-logical underpinning of high-stakes testing—the use of incentives, per se—is flawed. Even if we got the details right, the whole approach is likely to do far more harm than good.

Still, I'm tempted to reply that there are a lot of incompetent teachers out there. Don't we have to resort to stronger measures to make them improve?

Certainly it's true that not all teachers—or representatives of any profession, for that matter—are inspiring and impressive. But the relevant question is whether a "doing to" strategy is likely to be more effective at helping them improve than is a `working with" strategy. (A related question is which approach is likely to attract more talented people to the field of education.)

All of the research showing that rewards and punishments are at best ineffective, and more commonly counterproductive, challenges the assumption that people can be bribed or threatened into getting better at what they do. Granted, it's often hard to craft a feasible alternative for staff development, but that doesn't argue for persisting with a heavy-handed tactic that clearly doesn't work. Policy makers who deal with recalcitrant teachers—not unlike teachers who deal with

recalcitrant students, by the way—yearn for a solution that's both easy and effective. Unfortunately, when they can't have both, they often settle for easy.

Linda McNeil of Rice University points out that, paradoxically, the test-driven instruction that takes place as a result of accountability-based reforms may reinforce what the *worst* instructors have been doing. "Under a prescriptive system of curriculum, student testing, and teacher assessment," she observes, "the weakest teachers were given a system to which they could readily conform."

Yet I know I've read in the newspaper about states and districts that have used what you call "heavy-handed" tactics and seem to achieve some success.

Here are four reasons you should be very careful about drawing lessons from those stories:

First, high-stakes testing and other "doing to" tactics have sometimes been tried right around the same time that other, more reliable strategies were being implemented. In Texas, for example, many observers have argued that, to the extent there has been any improvement in student performance it is "largely the result not of the tests, but of smaller class sizes, rising overall spending on education and a court-ordered equalization of resources between schools serving the rich and the poor.") In fact, it's entirely possible that positive results in such a scenario could have been even more impressive in the *absence* of high-stakes testing.

Second, claims of miraculous improvements often turn out to offer more hype than hope. One illustration: A closer look at the allegedly amazing progress made by San Francisco public schools in the mid-1990s revealed that thousands of students who speak very little English had been excluded from testing right around the time the city's scores started to rise.

Third, we need to track results for a while. Researchers have found a predictable pattern playing itself out in state after state. When tests are first administered, the scores are distressingly low. (And the headlines read: Our schools are failing! Our students are ignorant!) After a year or two, the scores begin to rise as students and teachers get used to the test. (And the headlines read: Our schools are improving! Tougher standards are working!) Then the scores level off or begin to drop—or, if a new test is substituted for the original one, even plummet. (We've grown complacent! Even *tougher* standards are needed!) Politicians and journalists assume they are watching a rise and fall in the quality of instruction, despite the fact that this familiar cycle is largely an artifact of the testing itself.

Fourth, and perhaps most important, keep in mind that claims of higher achievement are almost always based on the very test that was administered to "raise standards." As anyone with even a smattering of knowledge about educational measurement will tell you, a given test cannot be used as a lever (that is, as part of a high-stakes program that says "Make these scores go up, or else") *and* as a measure of the success of that program. You're not getting a valid picture of learning; you're getting a reflection of students having been drilled relentlessly to beat this particular test.

In thinking about claims of improvement, never forget that standardized tests in general are quite limited. Anyone who argues that an accountability program (or any new policy, for that matter) has been successful ought to be asked what exactly is meant by `successful." If, as is often the case, the claim rests upon nothing more than higher test scores, we would do well to reply, "Given what we know about these tests, you have yet to offer meaningful evidence of success." (This point is especially relevant when heavily scripted programs involving direct

instruction of low-level skills are justified as "effective" solely on the basis of short-term test gains.)

Speaking of meaningful evidence, I'm wondering whether there is any research specific to the effects of high-stakes testing.

Not much. That in itself is remarkable; it means that our children are, in effect, being used as involuntary subjects in a huge high-stakes experiment. But what's worse is that the limited evidence that *does* exist suggests that this approach isn't even successful on its own terms—that is, at promoting narrowly defined academic achievement. Historically speaking, high-stakes testing has "failed wherever it has been tried," according to Linda Darling-Hammond, professor at Stanford University. And in the mid-1990s, states *without* high-stakes exit exams actually showed more improvement on another standardized test, the eighth-grade National Assessment of Educational Progress (NAEP), than states with such graduation exams. Evidence from other countries is similarly discouraging.

Also relevant here are small-scale studies that look at how various approaches to school reform affect individual classrooms. Researchers at the University of Colorado asked a group of fourth-grade teachers to teach a specific task. About half the teachers were told that when they were finished, their students must "perform up to standards" and do well on a test. The other teachers were simply invited to "facilitate the children's learning." The result: Students in the "standards" classrooms did not learn the task as well as those in the other group.

Really? Why would teachers who had their attention focused on bringing up the scores end up with students whose scores were lower?

Ask a roomful of teachers to speculate on why that happened, and you'll get a roomful of different answers, almost all of them plausible. Here's one clue: In a similar study conducted in upstate New York, teachers in the "standards" condition were observed while they taught. Essentially, they turned into drill sergeants, removing any opportunity for the students to play an active role in their own learning. When the teachers were controlled, in other words, they responded by becoming controlling. That makes it harder for real learning to take place.

Can we back up a moment here? You said a little while ago that the high-stakes approach is not even successful on its own terms"— that is, at promoting achievement. What other terms are there?

Even if it did boost achievement, you'd have to weigh that against the other things it does.

First, it *drives good teachers and principals out of the profession.* Teachers are already beginning to tire of the pressure, the skewed priorities, and the disrespectful treatment as they are forced to implement a curriculum largely determined by test manufacturers or state legislators. Some are talking about quitting—or at least avoiding the grade levels where tests are routinely administered, such as fourth grade. The most promising teacher candidates, too, may be reluctant

to begin a career that is increasingly centered on test results rather than on learning—or to work in a system that will try to manipulate them with rewards and punishments.

Similarly, it's becoming difficult to find qualified people who will agree to take a position as a principal. One middle school principal in Kentucky says he has watched his colleagues "disappear from the ranks. No one wants to blame it on [high-stakes testing programs], but from my perspective as a practicing principal, many of them made it clear they weren't going to put up with unreasonable demands." Because those who are leaving include some of the best teachers and administrators, the paradoxical result, once again, is that the "tougher standards" movement has the effect of lowering standards.

Second, even if they stay, educators may become *defensive and competitive.* In a high-stakes environment, teachers and principals understandably may feel the need to prove that low scores were not their fault. Moreover, it may set them against one another:

> A state with which we are familiar adopted a program that based high school mathematics teachers' annual raises on gains in their students' achievement scores. The next year some of the top teachers in the state resigned in disgust. Those who remained entered into intense competition with one another, which disrupted school programs and caused morale to drop throughout the state. (Among other things, some math teach-ers demanded that their schools restrict extracurricular activities, cancel school assemblies, and abolish our-of-school trips that might interfere with their instructional efforts.) The fol-lowing year the incentive program was dropped.

Third, high-stakes testing has led to widespread *cheating.* Educators in state after state, pressured to raise test scores, have been caught coaching students inappropriately during tests or altering answer sheets afterward. Reports of such behavior always elicit condemnation of the individuals involved but rarely lead people to rethink the pressures attendant on high-stakes testing. Other dubious tactics, meanwhile, are likely to be ignored entirely, such as providing extensive support for students who are right on the border of being able to pass the tests and slighting everyone else. Some educators may even force low-achieving students to repeat a grade, not because this is likely to be in the students' best interest (which it almost never is) but because it's assumed that they'll do better on the exam the following year and in the meantime won't bring down the average of the current pool of test-takers. Low-scoring students may also be designated as "special needs" to exempt them from the tests, thereby bolstering the school's overall standing.

Fourth, high-stakes testing *may turn teachers against students. A* superintendent in Florida observed that "when a low-performing child walks into a classroom, instead of being seen as a challenge, or an opportunity for improvement, for the first time since I've been in education, teachers are seeing [him or her] as a liability " Needless to say, if educators "resent children who are likely, for one reason or another, to perform poorly, they cannot establish the nurturing relationship with those children that will enable the children to trust them.

Fifth, it may contribute to *overspecialization.* In Ohio, the pressure to boost proficiency test scores has contributed to changes in how teachers of children from age nine to fourteen are certified by the state, forcing them to specialize in only two content areas, such as math and science. This means that the kind of departmentalization that has created such a fragmented educational experience in high school may now happen, thanks to testing pressures, as early as fourth grade. (Departmentalization, in turn, tends to support other problematic practices, such as the use of letter grades and the segregation of students by alleged ability.)

Sixth, it *narrows the conversation about education.* The more that scores are emphasized, the less discussion there is about the proper goals of schooling and the more educators are reduced to finding the most efficient means for what has become the de facto goal: doing better on tests. Furthermore, there is less inclination to use (or develop) alternative assessments. As long as a school or teacher has adequate test scores, what happens in the classroom is irrelevant"; poor test scores, meanwhile, are viewed as indicators that change is needed, "no matter what happens in the classroom."

Finally, there's the big one: the most predictable consequence of high-stakes testing, which is being noted with increasing bitterness by teachers all over the country but is rarely understood by those outside the classroom.

And that is . . . ?

High-stakes testing has radically altered the kind of instruction that is offered in American schools, to the point that "teaching to the test" has become a prominent part of the nation's educational landscape. Teachers often feel obliged to set aside other subjects for days, weeks, or (particularly in schools serving low-income students) even months at a time in order to devote themselves to boosting students' test scores. Indeed, both the content and the format of instruction are affected; the test essentially *becomes* the curriculum. For example, when students will be judged on the basis of a multiple-choice test, teachers may use multiple-choice exercises and in-class tests beforehand. This has aptly been called the "dumbing down" of instruction, although curiously not by the conservative critics with whom that phrase is normally associated.

More strikingly, teachers will dispense with poetry and focus on prose, breeze through the Depression and linger on the Cold War, cut back on social studies to make room for more math—all depending on what they think will be emphasized on the tests. They may even place all instruction on hold and spend time administering and reviewing practice tests. The implications for the quality of teaching are not difficult to imagine, particularly if better scores on high-stakes exams are likely to result more from memorizing math facts and algorithms, for example, than from understanding concepts. As two researchers put it, "The controlling, `top-down' push for higher standards may actually produce a lower quality of education, precisely because its tactics constrict the means by which teachers most successfully inspire students' engagement in learning, and commitment to achieve."

Teachers across the country struggle with variations of this dilemma, worrying about their jobs as well as the short-term price their students may have to pay for more authentic learning. The choices are grim: Either the teachers capitulate, or they struggle courageously to resist this, or they find another career. "Everywhere we turned," one group of educators reported, "we heard stories of teachers who were being told, in the name of 'raising standards,' that they could no longer teach reading using the best of children's literature but instead must fill their classrooms and their days with worksheets, exercises, and drills." The result in any given classroom was that "children who had been excited about books, reading with each other, and talking to each other were now struggling to categorize lists of words."

Even in classes less noticeably ravaged by the imperatives of test preparation, there are hidden costs—opportunities missed, intellectual roads not taken. For one thing, teachers are less likely to work together in teams. For another, within each classroom "the most engaging questions kids bring up spontaneously—`teachable moments'—become annoyances."

Excitement about learning pulls in one direction; covering the material that will be on the test pulls in the other. Thoughtful discussions about current events are especially likely to be discarded because what's in today's *paper* won't be on the exam. Furthermore, it is far more difficult for teachers to attend to children's social and moral development—holding class meetings, building a sense of community, allowing time for creative play, developing conflict-resolution skills, and so on—when the only thing that matters is scores on tests that, of course, measure none of these things. Indeed, there is anecdotal evidence that a greater emphasis on heavy-handed discipline to enforce order may be one more consequence of the imperative for test preparation.

These disturbing changes can take place whenever people's *attention* is drawn to test scores. But if bonuses for high scores are dangled in front of teachers or schools—or punitive "consequences" are threatened for low scores—the chances are far greater that a meaningful curriculum will be elbowed out to make room for test-oriented instruction. And this is most likely to happen in schools that serve low-income students

To talk about the kind of teaching that takes place in the name of raising scores is to talk about the kind of teaching that is abandoned. First to be sacrificed in a school or district where rewards or punishments attend the results of such testing is a more vibrant, integrated, active, "student-centered" kind of instruction. (Arguably, the alternative to a student-centered classroom today is not one that is teacher-centered but one that is legislature-centered.) The more prominent and relevant the tests become, the more difficult it is for teachers to invite students on an intellectual adventure, to help them acquire the ability and desire to solve realistic problems in a thoughtful way. One example can stand in for thousands:

Kathy Greeley, a Cambridge, Massachusetts, middle school teacher, had devised a remarkable unit in which every student selected an activity that he or she cared about and then proceeded to become an expert in it. Each subject, from baking to ballet, was researched intensively, described in a detailed report, and taught to the rest of the class. The idea was to hone researching and writing skills, but also to help each stu-dent feel like an expert in something and to heighten everyone's appreciation for the craft involved in activities they may not have thought much about. In short, it was the kind of academic experience that people look back on years later as a highlight of their time in school. But now her students will not have the chance: "Because we have so much content material to cover, I don't have the time to do it," she says ruefully. "I mean, I've got to do the Industrial Revolution because it's going to be on the test."

But surely not all states have tests that are as fact-based as the one that's apparently being given where this teacher lives.

Actually, the tests in most other states are worse! So if exams like the Massachusetts Comprehensive Assessment System (MCAS), which purportedly requires problem solving and higher-level understanding, have the effect of squeezing out some of the best teaching, imagine how dire the situation is in states where the tests are truly appalling.

Indeed, more and more teachers around the country feel compelled not only to teach testable facts in test-like fashion but to impart advice about test-taking, per se. This is not only an egregious waste of time but educationally harmful to the extent that students begin to generalize such strategies—for example, adopting the habit of skimming a book, looking for facts they might be asked about on a test, rather than thinking deeply about and responding to

what they are reading. But if clever strategies (for example, skipping to the questions first, then going back to the passage to find the answers) *are* effective, this means that a high test score is partly just a function of good test-taking skills. If students' scores can indeed be raised by teaching them tricks or by cramming them full of carefully chosen information, this should be seen not as an endorsement of such methods but as a devastating revelation about how little we have to learn from the results of these tests.

Linda Darling-Hammond offers this analogy: Suppose it has been decided that hospital standards must be raised, so all patients must now have their temperatures taken on a regular basis. Shortly before the thermometers are inserted, doctors administer huge doses of aspirin and cold drinks. Remarkably, then, it turns out that no one is running a fever! The quality of hospital care is at an all-time high! What is really going on, of course, is completely different from providing good health care and assessing it accurately—just as teaching to the test is completely different from providing good instruction and assessing it accurately.

Is that really a good analogy? Those doctors are simply cheating.

Right. And that's exactly why the analogy is apt: Teaching to the test could be described a "legal cheating." However, unlike the kind of cheating that is widely condemned (giving kids the answers) or the kind that would be condemned if it were publicized (flunking potentially low-scoring students or shuffling them off to special ed.), drilling students so they'll do well on the test even if they're not really learning much of value is generally accepted. In many areas, it is expected; indeed, it is even *demanded* by people who think that only test scores matter—or people who are bullied into acting as though they thought that was true.

Remember: If one district or school outscores another, a hefty part of that difference is probably due to socioeconomic factors and is therefore pedagogically meaningless. But even if we focus on a single district or school, in effect holding those factors constant, improvement on standardized tests over time may be worse than meaningless; it may be reason for concern. Time spent preparing students to succeed on such tests is time that could have been spent helping them become critical, creative, curious thinkers.

But do these imperatives have to be mutually exclusive?

Even for us to acknowledge that they are conceptually distinct, that they *could* lead to different practices, would represent a big improvement over the tendency to conflate the two—a tendency reflected in speeches offered by politicians, reports issued by business groups, and articles published in the popular press, all of which talk about "excellence" and "raising the bar," "tougher standards" and "higher expectations," and clearly mean nothing more than higher scores on standardized tests.

In practice, higher scores do not necessarily signal higher-quality learning. At a recent gathering of educational measurement experts (sponsored by the National Science Foundation and the RAND Corporation), agreement emerged that "in states where test scores are rising, the improvements may have nothing to do with whether schools have upgraded their teaching and curricula" but instead reflect "students' and teachers' increased familiarity with the state assessments" and "improved test-taking skills unrelated to the curriculum."

But we can go further than this. Just as we've seen that high scores and deeper thinking tend to be inversely related, so it is that teaching geared toward higher scores and teaching geared

toward higher quality learning can often pull in opposite directions. It's naive to believe that teachers can continue providing the best kind of instruction while remaining confident that their students will do just fine on standardized tests. If a test requires coverage of a great deal of material—however superficially—then exploring a few things deeply will be poor preparation for the test even though it may be far more effective for achieving various intellectual goals. This is especially true in science and social studies, where the best way to teach (according to a growing consensus among educators in those fields) is diametrically opposed to the best way to raise test scores, which may involve textbooks, lectures, and worksheets to promote memorization of dates and definitions. The former is about discovery; the latter is about coverage. Thus, "schools that frown on teaching to tests might be singled out as `underperforming"—penalized for doing what is best for kids.

Linda McNeil's description of the choice faced by Texas educators will be instantly and painfully familiar to teachers across the country—but may come as news to some parents, school board members, politicians, and reporters:

> The myth of the proficiencies [that is, the standards] was that because they were aimed at minimum skills, they would change only the weakest teaching. The "good" teachers would as a matter of course "already be covering" this material and so would not have to make adjustments. In fact, the transformation of the curriculum into received knowledge, to be assessed by students' selection of one answer among four provided on a computer-scored test, undermined both the quality and quantity that "good teachers" could present to their students.... [Thus,] teachers faced serious ethical dilemmas. They could teach to the proficiencies and assure high test scores for their students. Or they could teach the curricula they had been developing (and wanted to continue to develop) and teach not only a richer subject matter but also one that was aimed at students' understanding and their long-term learning, not the short-term goals inherent in the district testing of memorized fragments. This was not an easy choice.

Incredibly, states like Texas, North Carolina, and Virginia, where the tests are among the worst and the use of rewards and punishment to mandate a curriculum centered on those tests is particularly heavy-handed, have been held up as models for other states to follow. Such is the educational climate of our times: you can't be too tough, you can't test too much, and there's no policy so ludicrous that you can't get away with it—even brag about it—as long as you remember to recite the magical words "accountability" and "tougher standards."

Again, though, even in states where the standards are less objectionable, the one-size-fits-all testing systems have approx-imately the same effect on quality curriculum that a noose has on breathing. Today's mail brings a report of an innovative medical mentorship program at a middle school in New Rochelle, New York, where eighth graders follow doctors through the day and write rigorous research papers. "Like dozens of other programs statewide that use assessments other than standardized exams," this one "doesn't comply with the mite's new educational standards" and is slated for extinction.

Part of the problem, it should be said, takes us back to the Assumptions about quantification that lie behind the push for tougher standards: To talk about what happens in schools as moving forward or backward in specifiable degrees, to make a fetish of "specific, measurable goals," is not only simplistic insofar as it fails to capture what is really going on; it is destructive insofar as it changes what is going on for the worse. Once teachers and students are compelled to focus only on what can be reduced to numbers, such as how many grammatical errors are present in a composition or how many mathematical algorithms have been committed to memory, the

process of thinking has been severely compromised and the best programs and classes can't survive.

The ultimate result of this sensibility is, as critics of one state's reform efforts observed, that "what can be measured reliably and validly becomes what is important to know." This is the opposite of what would seem to be the sensible way to formulate an education policy—namely, to begin by agreeing on some broad outlines for what students ought to know and be able to do, and then address the question of assessment. These days, it's just the opposite: the tail of testing is wagging the educational dog.