Standardization Versus Standards

In the name of objectivity and science — two worthy ideas — the testing enterprise has led teachers and parents to distrust their own ability to see and observe their own children, Ms. Meier points out. What we need are assessments — with low or high stakes — that place authority in the hands of people who actually know the students and that make sure that the community, the family, and the student have ways to challenge such judgments.

BY DEBORAH MEIER

PROPONENTS OF THE current so-called standards-based reform, including state and national government leaders, business leaders, and editors of most of our leading newspapers, claim that the way to restore trust to public education is through objective tests. They argue that it is possible to design tests that can stand the weight of accountability, determine high-stakes decisions, direct good teaching, and tell where everyone stands in relation to everyone else — and define what it means to be well educated.

The search for such a "good" test — one that gets around the difficulties posed by the norm-referenced ones that have dominated the last century and can drive school reform — keeps us tied to a false hope, however well intentioned.

One can see the appeal, however. Reformers of all stripes have always hoped there was a way to do this. Design a test with norms based on what people should be able to do, not just the range of how they currently perform; it would be more like a driver's test. Wouldn't it make all our jobs easier if we could find a way to measure everyone against an absolute standard of what it means to be well educated? Wouldn't this help direct the changes we want in schools (and society) and focus our attention on the acknowledged weak spots? Even if people didn't at first agree on our definition of the standard, wouldn't most people go along simply out of the desire to do well? The

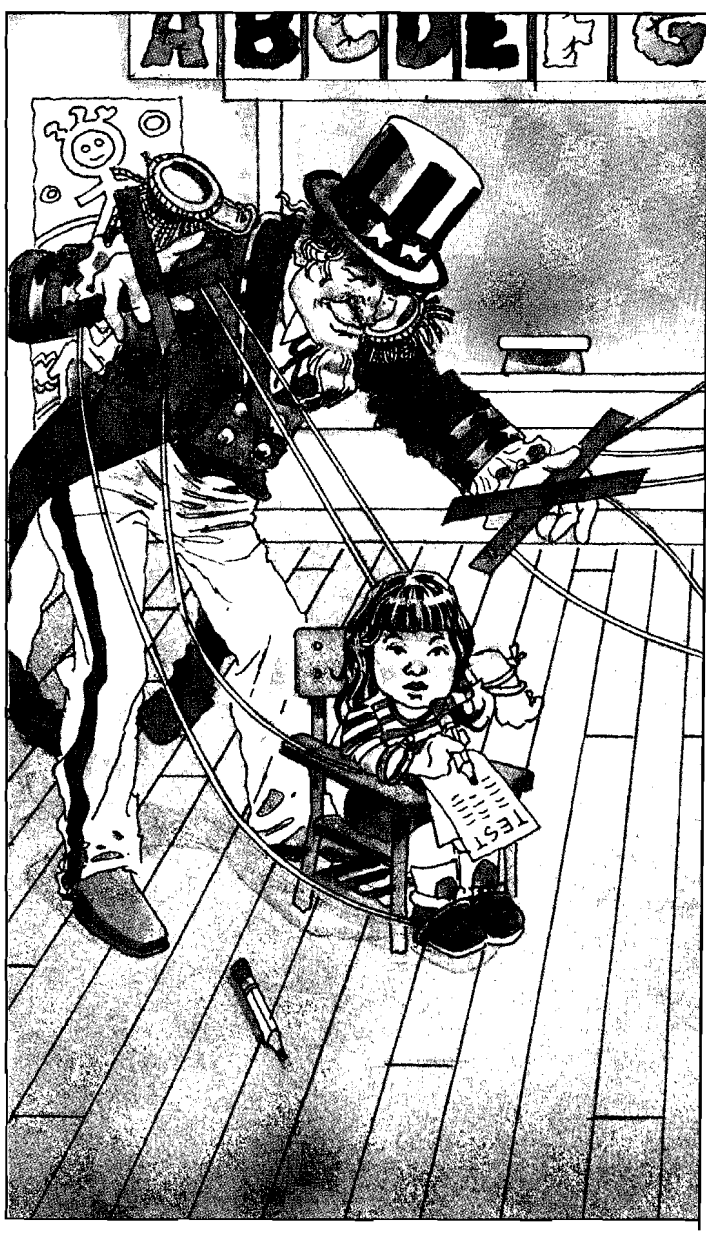


Illustration: Jim Hummel

DEBORAH MEIER is the MacArthur Award-winning founder of Central Park East School in East Harlem and of Mission Hill School in Boston. *This article is adapted from her new book,* In Schools We Trust: Creating Communities of Learning in an Era of Testing and Standardization *(Beacon Press, 2002). The book is available in bookstores or by contacting Beacon Press at www.beacon.org, ph. 800/225-3362.*

THE PURPOSE OF THIS NEW WAVE OF TESTING IS NOT, REMEMBER, TO OBTAIN MORE DATA. **THE PURPOSE IS TO CHANGE THE SCHOOLS.**

test would do the convincing. That's what standards-based reform is about — making change happen, raising our sights.

The purpose of this new wave of testing is not, remember, to obtain more data. The purpose is to change the schools. We already have more standardized, objective, and centrally collected information about our schools than any country on earth. We have test scores of every sort, at every age level, broken down every which way you can imagine — by race, class, gender, geography, and more — plus data on attendance and dropout rates, much of which goes back half a century or more. (For example, we've known for decades that no neighborhood high school in the Bronx graduates more than 30% of its incoming ninth-graders.) But the problem is that such measures, while they spot where there's trouble, don't actually do away with the trouble. Nonetheless, that seems to be the new idea: testing *as* reform, not *for* reform.

The popular new drive to hold schools and school reform accountable by means of test scores has many attractions. It's built around the idea that the villains are mostly low expectations and a failure of will. Since both are indubitably factors in failure — and less onerous to tackle than poverty, for example — this notion eliminates victimology. And it keeps us focused. Ordinary citizens, including parents and teachers, are aware of how often local parent councils, teacher unions, principals, and local school boards have abused their powers — here's a way to catch them. No more excuses. The more objective the "standards," the more distant and scientific the results; the more universal the population tested, the less negotiable the consequences and the less room for argument, excuses, flexibility, bias, and compromise.

In a society in which adults often feel helpless to control their students or their children, even to know them, this approach has additional blessings. It appears to avoid

the issue of trusting anyone: one's children, their teachers, their schools — or even oneself. It is, we are told, also more like the merciless but efficient and effective marketplace — with test scores standing in for the bottom line. And for this reason it also appeals to those who have the most reason to distrust our schools: urban minority families and those inclined to be suspicious of any public institution. Finally, we have a tool with teeth, one that offers both clear and universal goals and direct observable consequences for not meeting them.

The idea of holding schools accountable for test scores has its attractions, fits aspects of the national mood, and adheres to a long-standing American tradition of turning to standardized testing as the cure for our ills. The trouble is, as we keep relearning generation after generation, it contradicts what we know about how human beings learn and what tests can and cannot do. That a standardized one-size-fits-all test could be invented and imposed by the state, that teachers could unashamedly teach to such a test, that all students could theoretically succeed at this test, and that it could be true to any form of serious intellectual or technical psychometric standards is just plain impossible. And the idea that such an instrument should define our necessarily varied and at times conflicting definitions of being well educated is — worse still — undesirable.

THE SO-CALLED NEW TEST

In the late Nineties, states sought to impose by way of tests newly designed state curricula — keyed to, or in some cases interchangeable with, a set of agreed-upon standards. This development made more obvious the essential contradiction between a testing system designed to be secret and normed to fit a bell curve and the purposes of the new reform agenda, in which everyone was expected to achieve success. The answer: a new kind of test, one that could be directly taught to, didn't require as much secrecy regarding content, and above all no longer required scores that distributed students along a predetermined curve. Everyone is urged to adopt these new tests — although rank ordering and percentile scores are still used. These tests are intended to show whether teachers and students are doing their prescribed jobs: teachers teaching to the test and students learning what's on them. It's called curriculum and test alignment. A number of states developed variants of this new sort of test — the Massachusetts Comprehensive Assessment System (MCAS), the Regents Examinations in New York, the Texas Academic Assessment System (TAAS), and the Standards of Learning (SOL) in Virginia, to name a few.

From the viewpoint of the test-taker, these are very similar to the old tests, though generally they are much longer. From the viewpoint of the teacher, the big difference is that these tests can be taught to openly. From the viewpoint of the state, the scores are set not by the test-makers but by political officials in state departments of education. One might describe these as politically rather than technically normed tests. For example, the weighting of subsections — how much each counts — and thus the actual scores and what score constitutes failure, what constitutes needs improvement, what constitutes proficient — are in many states not decided until after the results are in and state officials can estimate the impact of their decisions. (But in all states pretests give a pretty accurate estimate.) The meaning of a score on these new tests rests not with the neutral bell curve but with judgments made by some politically appointed body — ideally in collaboration with educational experts.

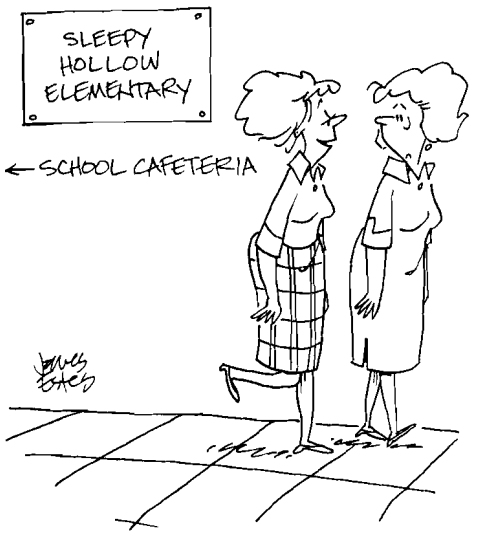
The new tests are more like the ones teachers or academic departments have long been accustomed to giving at term's end — covering what they think were the key elements of their courses. When they are the ones to set the scores, teachers too are influenced by political factors — who will blame them if the scores are too low, will they be believed if they are too high, what's the school's attitude toward marking on a curve? The technology is not necessarily dissimilar — teachers often use multiple-choice exams, for example. But unlike the designers of the new state tests, classroom teachers and local administrators are folks close to the action, "interested parties" who can modify their exams and scores based on their best judgment and who are aware of what actually is happening in their classrooms and schools. Of course, their very closeness is the reason why, in today's climate, teachers are distrusted.

How different are these new tests to design than the traditional norm-referenced tests? Largely, the answer is, not a lot — except that the absence of the much-maligned bell curve complicates deciding what items to include and how to set expectations, scores, and cutoffs. Creating these tests begins the same way as for any standardized test. Hundreds of teachers and expert academicians, under the direction of the (politically established) state education department, develop their wish lists of things they believe all students should know, appreciate, understand, and be able to do at particular ages or grades; ideally, these wish lists are tempered by experience.

For example, one might wish all third-graders could read the Harry Potter books — but is this goal reasonable? What about *To Kill a Mockingbird?* What about Shake-

speare? Reading the California art standards for kindergarten, one is inclined to think that test-makers had in mind the scope and sequence of a postdoctoral program in the arts. Could they possibly have had 5-year-olds in mind when they wrote that "students will research art genres (e.g., landscapes, seascapes, portraits), name an artist who worked in the genre, describe the artist's work, and then create an artwork that reflects the genre" or that "students will talk about a work of art, telling what they think the artist is saying, and give reasons for their responses, using art terms (line, color, shape)" or that students will "compare and contrast a Renaissance landscape and a landscape by Richard Diebenkorn"? (Actually, the last of these examples came from the first-grade standards.) In case you are curious, not only are similar requirements set for dance — "compare and contrast American square dances and English contra dancing," for example — but the same amazing expectations are repeated in every other subject discipline. And California is not notably different from other states, nor are the arts standards any more humorous than those in history, math, literature, and science. When I sat on the New York Regents advisory board, I ran across the following in health education for 12-year-olds: students will demonstrate that they can cope with death and dying, as well as losing a friend. Why not?

Decisions regarding how to go from such pretentious wishes to actual items on a test are difficult, since they can't be based on how things would sort themselves out on a bell curve or any other predetermined ranking (which would quickly cure test designers of such nonsense). In the absence of such a curve, decisions can be made that almost all children are appallingly lacking in artistic talent or coping skills and that they need earlier and more intensive remediation. Drill and practice in coping with death or identifying landscape genres?



"I grew up in the glide path of a big-city airport — so the decibel level in there doesn't bother me."

However the decisions are made, the items will now produce a detailed scope and sequence of facts and skills to be taught from kindergarten through 12th grade. From now on, the field is level, so proponents would argue: everyone knows what it is that might be on the test. Vague goals like "weighs evidence" or "writes with style" are hard to score objectively — and harder to teach to. Thus they are eliminated. The new lists are often long. Robert Marzano and John Kendall of the McREL education institute figured out that covering all the standards on the average state list would take nine more years of schooling. But no one wants his or her favorite items eliminated from the curriculum framework, since it is probable that only the stuff that makes it onto the test will ever be taught.

It is important to note that, because the idea in many states is to at least appear more and more demanding, there is no obvious way to agree upon the reference base, as there is with traditional norm-based standardized tests. "But kids that age can't do that" and "teachers can't cover all that" may meet the response "but they should be able to." I believe that this is more than just an apparent nuisance: it is at the heart of why these tests *cannot* deliver what they promise. There are also some knotty content decisions that make such tests sink or swim politically: how (and whether) to teach about evolution, the Civil War, the labor movement, Reagan's place in history, the causes of World War I, or — as the state of Virginia (as I write this) is now finding out — what to say about the role of the Turks in the Armenian "genocide," not to mention even what to call it. Decisions on these issues must now be made at the highest levels, and they must be given teeth so that they can be enforced in the form of tests.

In fact, although it's easiest to see such controversies in the fields of social science and history, they abound as well in math, science, and literature. California's efforts to implement such a test were derailed a decade ago by the choice of certain multicultural texts, as well as by writing assignments that asked students to write about personal experiences, and both California and Massachusetts are embroiled in wars over what math students should know and when. Of course, no one is planning to add nine years to the schooling of every child. And in real life, good sense takes over, and schools actually prepare students only for a sufficient amount of the material that they discover, over time, is actually likely to be on the tests and is necessary to achieve a passing score and so ensure that the schools look good compared with the competition.

The critical decisions involve the actual selection of which items from that long list to include on a particular test, as well as the wording of the questions and the possible alternatives offered. Not everything in the curriculum framework can make it into one test! What kind of "distractors" — alternative choices that are wrong — should be included and how to decide? Posed one way, the question will be an easy item; posed another, it will be hard. For example, "Was Lincoln the first or 16th President of the United States?" is easy and important to know. "Was Lincoln the 13th or the 16th President of the United States?" is hard and is arguably not important to know. But both items may be used to enforce a standard that asserts that students should know when Lincoln was President. For this part of the work of test development, the process is much the same as it was for the old standardized tests, involving both sample pretesting and statistical analysis — but again with a difference. For the old tests, the deciding factor was whether the scores produced were sufficiently and appropriately spread out; now that is not necessary.

After the pretesting, another difference between these new tests and traditional psychometric tests emerges. Since there's no need to tweak the results to fit a rank-ordered curve, the issue now is simply what to call the scores. When the first student took the newly minted MCAS, the Massachusetts Department of Education was free to decide that 80% of all students would be labeled less than proficient and so be judged to have failed as readers and that 0% demonstrated advanced status as writers. Since Massachusetts ranks high in language arts on all nationally normed tests, including the SAT and the National Assessment of Educational Progress (NAEP), the decision may have seemed odd. In fact, the ensuing storm caused the department to lower the bar — amid protests that this was dumbing down the test — so that only 80% of *urban* students would fail. Opposition continued to increase, and by the fourth year the state department fielded a test in which fewer than half of all urban students failed.

One celebrates and weeps simultaneously at the enormous distraction involved, at the waste of time and energy in pursuit of the wrong goals.

Given the above oddity, it's not surprising that a test advertised to test "standards" becomes whatever is needed: a minimum competency test in some states (as in Texas and North Carolina) or a "tough" test (as in Massachusetts, Virginia, and New York, though now a student can eventually pass the MCAS with just 33% of the answers correct on the math test). Richard Rothstein reports in the *New York Times* that in the spring of 2000, 98% of Ohio students passed their high school graduation test, whereas less than half passed their test in California. And even fewer would have passed if California had stuck with the educators' recommendations rather than those of Delaine Eastin, the state commissioner.

Anomalies of all sorts abound. Only 28% of eighth-graders were scored as proficient on the Massachusetts science exam, although their scores on international tests show them outranking every nation except Singapore. Conversely, North Carolina's state test showed 68% of students proficient in math, whereas only 20% were judged proficient on a national science exam. The NAEP does not fare much better. Only 2% of high school seniors were labeled advanced on the NAEP math test, but twice that number alone pass Advanced Placement exams in math each year, and about 10% score above 600 on the SAT math subtest. Who is right? Who is wrong? These absurdities result from trying to adapt a technology that was never designed for such purposes.

In addition, such tests face a whole host of related problems that stem from the central fact that they have no basic reference point except political judgment. Equating tests — a technical term for comparing scores on different tests or on different forms of a test that change from year to year — is another once-minor headache that these new tests have compounded. For example, Massachusetts has, to its credit, decided to make most items public each year; in other states, the frameworks have changed frequently. In either case, new tests are needed. So is a score of 72 on one test the same, higher, or lower than a score of 68 a year later on a new test? Discussing test rescoring in Texas, psychometrician Daniel Koretz acknowledged in *Education Week* that equating posed serious problems in the context of standards-based testing. Texas officials claimed that their 2001 test was harder than their 2000 test, that lower raw scores didn't mean lower performance — so they had added credit. The Massachusetts fourth-grade English test was made easier in the third year in response to complaints that the reading passages were almost all on a sixth- through 10th-grade level of difficulty. When challenged regarding how scores should be compared from the second to the third year, the state department reassured the public that, while the questions were easier, the students now needed more right answers to get the same score. A similar problem arose in New York City when sixth-grade scores were unaccountably much higher one year, owing — the test-makers said and New York City officials denied — to equating. Of course, there were substantial consequences for promotional policies.

What is thus strikingly different about these new variants is not the tests themselves but the chutzpah of those

who design and use them for high-stakes purposes despite these unresolved issues. The designers of the old tests, who expected their tests to last a decade or longer, frankly claimed that teaching to them was unfair and invalidated the meaning of the scores. They argued that the items had not been selected for that purpose. The careful and fairly modest claims for when and how the tests should be used and the high measurement error involved in any single score stand in stark contrast to current claims for these new, less rigorously designed tests.

The biggest differences between the old and the new state-designed tests is that the new tests are put together much faster, require less technical validation and fewer reliability checks, are much longer, include more detailed factual questions, and are used for more high-stakes purposes. In addition, the scores are no longer a mere artifact of the bell curve but are instead a mere artifact of the judgment of state commissioners.

Each of these differences ought to be controversial. Yet they rarely are. And there are more differences. For example, makers of the traditional psychometric tests claimed that tests for elementary school pupils were actually less reliable if they lasted too long — the scores would be influenced by the sheer exhaustion of the students. An hour was viewed as the limit of technical reliability for children under age 10. But tests that do not meet such criteria are routine for children who are 7 and 8 years old these days. Test-makers used to insist that the degree of measurement error (which was routinely made available to schools) precluded using scores for any high-stakes decisions. A score of 4.5 on a test did not mean that the student was reading like a fourth-grader in the fifth month of the year (which is how the numbers are translated into English). In all likelihood the true score was somewhere between 3.9 and 4.9 — and possibly even higher or lower. Yet diplomas now hang on much finer lines of demarcation. Psychometricians haven't changed their minds, but the tests are now being used to do what psychometricians once claimed was undoable.

"I considered home schooling, but then I realized they'd be home all day."

The test-makers agree that cities and states often use and abuse their tests. They themselves make modest claims, if asked, for what a test can tell us about individuals or schools. For example, "I am led to conclude," says Robert Linn, perhaps the preeminent leader in the field, "that the unintended negative effects of high-stakes accountability uses often outweigh the intended positive effects." But such statements carry little political clout, if they are noticed at all. The technical manuals, with their careful disclaimers, that accompanied such tests when I began teaching are no longer seen by schoolpeople.

THE IMPACT ON SCHOOLING

While this new breed of tests is remarkably similar to the old one, we are no longer warned against teaching to the test. In fact, state officials demand that we do so. The same publishers who make many of these new tests now publish coaching materials for their tests. If something is not likely to be on the test, the official word is, don't teach it. School officials in some states even argue that children's regular classroom grades should not be substantially different from their state test scores. In Boston, this wisdom was the basis of an explicit directive from the superintendent's office to all school personnel. Thus test scores and class grades do not become two different ways to measure progress but two ways to record the same test scores!

Because the tests now claim to measure exactly what should be taught, it is far easier (for better or worse) to script teaching down to a lesson for every day of the year, each corresponding to a set of potential test questions. Some districts mandate scripted lessons only for low-performing schools. This system makes it easier to standardize the textbooks to use (ones that conform to the state's frameworks) and the preparatory material to order (testing companies now have both hard copy and online material for virtually every state test). And it simplifies as well the design of teacher training.

Adopting such a system means that many a curriculum related to children's interests or contemporary or sponta-

neous events (a hurricane that just swept through town, the river that runs through the school's backyard, the arrival in town of an exhibit on the ancient Celts, the release of a great movie on World War II, or the attack on the World Trade Center) must be ignored — or at best noted only in passing — in order to cover the standardized test-driven fare. It's hard to justify spending whole months on any topic, much less one that might involve only one or two questions on the test — such as ancient China or the Holocaust. The 1999 MCAS test, for example, included one item on China — which required knowledge about the 13th-century Song Dynasty — and none on the Holocaust. Furthermore, unless tests are devised for all subject areas, everything not being tested — music, dance, the visual arts — is driven out of the curriculum.

THE OLD DISGUISED AS THE NEW

The majority of the states that have jumped on this new bandwagon still use the same standardized norm-referenced tests, but they now use them for this new and different purpose. Obviously impossible? State officials claim that it's reasonable to expect all students to be in the top half (or wherever the marker is set) of the distribution, even though, if the test-makers don't abandon their psychometric reputations entirely, that will lead only to a raising of the grade-level cutoff score sometime in the future. Oklahoma now has a law specifying that 90% of its third-graders should be on grade level on a currently normed test by 2007. If the superintendent is lucky, that is sufficiently far in the future that he or she will have moved on to another job somewhere else by then. Paul Vallas left his job in Chicago because he was still around when the sad news was announced that the high school scores were either unaffected by his reforms or actually going down, while the elementary scores — on a norm-referenced test used year after year after year — had gone up. (In fact, however, he went out claiming victory.) Both Oklahoma and Chicago are still using old-fashioned normed tests — with a twist.

The makers of the old normed tests have renamed their percentile scores with four levels, called advanced, proficient, needs improvement, and failed. But how they did this is unexplained. For example, on the norm-referenced Stanford Achievement Test (SAT 9), a student has to be in the 49th percentile to get a level II on the fourth-grade math test (i.e., to pass), whereas being in the 22nd percentile is required in language arts. These new names, labeled I through IV, have become the language of the standards movement and thus are commonly used on norm-referenced tests too. They were borrowed from the NAEP, the granddaddy of standards-based tests (originally NAEP made use of five levels of proficiency). The NAEP was designed by the U.S. Department of Education (ED) to gather longitudinal data based on small population samples. The NAEP's adoption of the four levels of proficiency, which is less than a decade old, is based strictly on judgment calls by a panel of ED-chosen experts with a reform agenda. The names and labels are whatever test-makers — including the publishers of the SAT 9, which is used in California, and the Iowa Tests of Basic Skills, which is used in Chicago, and their respective state authorities — choose to say they mean. Is there something Alice in Wonderland-ish about this?

IMPACT ON STUDENTS

In the meantime, the real-world consequences of these tests for a generation of youngsters, above all those already most vulnerable, hang in the balance. While critics claim that the high school diploma has become worthless, it continues to have a very exact monetary value — as we have been reminding children for years in all our "stay in school" advertising campaigns. The dollar cost of adopting these new graduation requirements will fall heavily upon communities of color. To deny increasing numbers of students a high school diploma will also mean that large numbers won't be able to enter our two- and four-year colleges, which will involve an even greater economic loss to the students, their families, and their communities. Whereas 70% of the seniors at Boston's famous Fenway High School failed the MCAS in 2000, before the high stakes went into effect, 90% went on to college, as they have for years, and did well there. Students at over 30 famous small high schools in New York City, such as Central Park East Secondary School, which have been sending 90% of their students on to successful college careers, are similarly endangered — unless those schools drop the very practices that produced such past success and focus on the test.

Test-mandated retention policies have similar chilling effects. Every time we hold a child back, we are substantially reducing the odds that that child will graduate at any time in the future. Once we hold a child back twice, the odds fall to less than 1%. Even before the standards movement attacked so-called social promotion, half of the young black men in America were at least one year over age when they reached eighth grade. What happens now?

The most significant impact of the new standardization is already evident in the increased dropout rate in state after state. In a detailed study of the "Texas miracle," Boston

University psychometrician Walter Haney documents how the very youngsters whom we recently wooed to stay in school are now being pushed out via tests. He notes that Texas continues to have the highest dropout rate in the nation. And dropout rates disguise the even larger number of students who "disappear" between sixth grade and 12th grade. Many supporters acknowledge the increased dropout rates but claim they represent a passing phase, the necessary price to be paid until the system and the students adjust. The leaders of the testing drive in Massachusetts are asking folks to wait and see. Headlines in the *Boston Globe* assert that, without pain, there can be no gain. These youngsters are, says the *Globe*, merely the necessary casualties of the war on behalf of high standards.

A state official in Massachusetts reassured legislators by noting that a student could get just 40% of the answers right and still pass. If one is measuring something important, getting 60% wrong and still passing is absurd. If one is measuring absurd things, however, it's another matter. It may be that the implicit denigration of the commonsense human judgment of the adults in young people's lives will be, in the long run, the greatest price paid in our current mania for high-stakes testing.

THE ALTERNATIVE TO STANDARDIZATION

The alternative to standardization is real standards. Standards in their genuine sense have always depended on the exercise of that suspicious quality of mind — trusting our fallible judgment — and training ourselves, as Jefferson recommended, to the better exercise of such judgment.

The best doctors know the danger of tests that seek to replace medical judgment. No diagnostic test stands by itself. And no diagnosis, no matter how uncontroversial, determines a good treatment plan. Treatment plans designed by HMO clerks or, for that matter, HMO doctors, far removed from patients, with access only to medical descriptions of patients' symptoms and copies of their test scores, are not what patients need. They need doctors with good medical training and good collegial and lay oversight, professionals accustomed to reviewing all the evidence. And second opinions must always be welcome. We are about to learn the same lessons in education.

To evaluate our local schools, we can collect evidence of various kinds in multiple forms, and we can bring in a range of external opinions — expert and lay — regarding the schools' reliability and validity. Debate, both local and national, is vital to the evaluation process. What we have to keep in the forefront is that data rarely speak for themselves. We must raise such questions as "What evidence is there that this is or isn't an important trend?" and "What do we know about how this plays out and what interventions work best?" After all, this kind of questioning is how we make judgments in most fields, including how we give doctoral candidates their Ph.D. degrees. It's how judges vote on movies, books, and the performance of Olympic gymnasts. It's even how we decide matters of life and death in our jury system. The jury handbook I received last year bragged about the fact that untrained citizens were entrusted to carefully weigh important matters. Only the most egregious self-interests are ruled out.

If we want to find out what teachers and parents can do to help a particular child's reading, we will have to seek to understand how that particular child is tackling reading tasks. Both traditional test scores and the relatively short interview we use at Mission Hill — consisting of a taping of a child's reading twice a year, followed by some standard open-ended questions — may be inadequate. We may need to obtain second and third opinions. No shame need be attached to the fact that we have only the most imprecise tools for making these kinds of assessments and that some are embedded in our daily interactions with the child and our close observation of the child at work in authentic settings. Two diagnosticians, be they teachers or doctors, may well disagree — even given the same set of x-rays — but it helps if they have other real-life symptoms to check their theories out on. The tasks of measuring and interpreting what is going on in a child's head call for trained judgment — our knowledge of what to listen for and how to recognize the array of misunderstandings that might lie behind a child's errors. But these are one-on-one tasks, and they are time-consuming. Good listening can be informed by science, although in the end it remains an art. The art of good teaching begins when we can answer the questions our students are really trying to ask us, if only they knew how to do so.

For those occasional gatekeeping purposes — quite a different matter — we can develop systems such as those described elsewhere that have been used at hundreds of middle schools and high schools over the past few decades for deciding when a child is ready to move on to the next level of schooling (systems that are also being challenged now by the imposition of high-stakes standardized tests). We have a history that demonstrates how such local performance-based systems work, and we have even had legislative proposals in various states to make these systems state policy. The systems vary; mostly they require schools and school districts to put together their own collection of standards, with a few spare common statewide indicators or tests.

All these systems combine careful expertise, public evidence, and eventual reliance on human judgment — not hidden behind tests but right out front. The doctor must explain why she is recommending one form of treatment or another and what the tradeoffs and side effects may be. She has to convince her patients, explain her reasoning, and discuss risks, not hide behind data as though the data spoke for themselves. Another doctor might disagree, might read the same sonogram or blood test differently based on other available evidence. Some patients might choose to change doctors. The same is true for educators. People often tell me that tests are part of real life, that kids need to be taught how to handle them. There's truth to this, and training in test taking is essential. But actually, far more often decisions are made not by test scores but by real-life judges in a format closer to the one we use at Mission Hill for portfolio reviews.

At Central Park East Secondary School, we used to combine our in-house judgments — our standards — with a wide range of external reviews. For example, each year we brought in a group of experts in one of the domains our students were required to pass muster on and had them assess our assessments. The experts' task was to critique us — the faculty — in an open and public forum. Their power was enormous, although there were no official sanctions attached to their findings.

THE PRICE PAID

What worries me most is that in the name of objectivity and science — two worthy ideas — the testing enterprise has led teachers and parents to distrust their own ability to see and observe their own children. In fact, objectivity and science start with such observation.

When parents and teachers no longer believe they can directly judge a child's reading ability, when they see the indirect evidence of tests as more credible, then I fear for the relationships between children and the adults they must depend on to grow up well. I worry, too, when children themselves can't tell us whether they are good readers until they see their scores. I know then that one of the goals of a good education — "know thyself" — has been lost. Cornel West says that Malcolm X added to this maxim: "to know thyself is painful." There are times that the "no gain without pain" message of the *Boston Globe*'s headline may apply; real self-knowledge is sometimes hard to come by. But avoiding it is not a solution.

We educators are paying the same price, as we anxiously wait each year for our students' test scores to be reported. We now depend on such scores to assess our own students and our own work The staggering jump in "achievement" of Massachusetts high school students between 2000 and 2001, for example, wasn't noticed by any of the system's teachers, students, or principals — at least not until the day the scores were released to the press.

Imagine the effect on a parent of a third-grader, beaming with pleasure at her son's apparent reading ability, when she discovers in a letter sent to her by the state that he really can't read. Imagine the reverse as well. The withering away of the expectation that human beings can and must make judgments, even on matters so intimate and close to home, has frightful side effects. And for the young, to be adrift in a world in which those who know them best are told that they do not know them at all undermines what growing up most requires: faith in adults and respect for their expertise. For a teacher who sees a student day in and day out to admit that she won't know how well he reads until the test score arrives is not good news. (And once we are convinced of the magic of test scores, how easy it is, by the mere act of setting "cut scores" wherever we wish, to convince the public at large that this or that percentage of children are or aren't doing well — depending on our purposes and agendas.)

Setting all children in the way of using their minds powerfully is well within our reach. Resorting to flawed standardized testing, whose only virtue seems to be its capacity to enable us to pretend we can rank everyone (or sort everyone) precisely and objectively, is both unnecessary and counterproductive to such ends. The development of a theory and practice of assessment that is consistent with the democratic demand for high achievement for all children is not impossible, and some of the ingredients for such a new approach already exist. What I hope I have demonstrated is that the current wave of standards-based tests is not the answer.

Tests are thermometers, not cures. At best, tests can take our temperature — sample where we are and hazard an educated guess at what a rise or fall in temperature might mean. Science simply won't solve these issues for us. As the old song goes, "We'll have to do it by ourselves." What we need are assessments — with low or high stakes — that place authority in the hands of people who actually know the students and that make sure that the community, the family, and the student have ways to challenge such judgments — asking questions, presenting competing forms of evidence, checking them out with a second opinion. We may find that old-fashioned standardized tests are one tool among many that will prove useful. We need, in short, standards in terms of both means and ends, not standardization. ■