Issues in Education Research

Problems and Possibilities

Ellen Condliffe Lagemann
Lee S. Shulman
1999

Jossey-Bass Publishers
San Francisco
CHAPTER FOUR

Needed

Thoughtful Research for Thoughtful Schools

Deborah W. Meier

Out walking one night I came upon a man circling under a lamppost. I asked him, "What did you lose?" "My watch," he said. "Are you sure you lost it here?"

"Actually," he admitted, "I lost it over there [he gestured into the dark], but the only light I can find is over here, so that's where I decided to look."

There has been too much research on schools built around the stuff that is easiest to look at—standardized, quantitative data that meet a particular set of psychometric criteria. What we need, however, is to understand what is off there in the dark. Keeping our eyes narrowly focused around the lamppost has helped us avoid asking aloud: "So what—and what for?"

We know that the so-called hard stuff—test scores, attendance data, dropout statistics—is not "hard" at all; it is soft and sloppy and highly accessible to our not-disinterested manipulations. Yet whenever I ask my colleagues why they continue to trust judgments based on such flimsy evidence, they chorus back: "It's the only stuff we have. Anyway," they add more softly, "our own voices and accounts are not considered credible."

But it may also be that reliance on existing forms of quantitative data serves a purpose that we dare not name.

Our curious dependency on a certain kind of "hard," or what Ted Chittenden, author and cognitive psychologist at the Educational Testing Service, calls "indirect," data can be partly explained by the all-too-human desire for simplicity and exactitude. The "direct" data—actual live children engaged in the thing itself, rather than subsamples of presumably related activities—require arguable interpretations and thus can produce complicated and ambiguous results. George Madaus, holder of the Boisi Chair of Education and Public Policy at Boston College, made the same point in his testimony to the Education Committee, Massachusetts House of Representatives (July 15, 1998). In criticizing a
new teacher test, he noted that it “flies in the face of a fundamental measurement principle . . . A direct measure of whatever skill or ability you are interested in is . . . preferable to any paper-and-pencil test. Further, even when a paper-and-pencil test is used, it should be validated against a direct indicator of the skill or ability.”

The public, we are told, will never buy that. Even at New York City’s Central Park East and now Boston’s Mission Hill schools, which I have been associated with for so many years, we have accepted this prejudice and used numbers (as well as narrative and anecdotal data) to report certain forms of direct observation.

It seems to comfort folks. And the education research community has helped to create the public’s willingness to settle for it. Rather than being an antidote to such oversimplification, experts and academics have fostered the dependency. If these well-informed persons use test scores almost exclusively to verify their judgments and treat such data as though they were the real thing, is it any wonder that the lay public believes that the true “ends” of education are best summarized by reference to quantitative measures of aptitude and achievement? In fact, it seems a scholarly quibble (or hopelessly romantic) to suggest that the ends of education do not nearly correspond to test scores.

But suppose the emperor wears no clothes. Suppose virtually all the data that schools and school systems are accustomed to use rest not only on outmoded ends but, worse, on a huge but useful deception. This chapter considers that possibility and suggests other, more valid approaches to the assessment of children’s academic performance.

I got a call one night from a wise friend who had played a critical role in the formation of a new elementary school designed exclusively for at-risk youngsters. “We’re starting our fourth year, and I’m ready to call it quits,” she said resignedly. “Our kids can’t read.” Given her past glowing accounts of this school, I was floored. “Didn’t you notice this before?” I asked.

“No,” she replied, “because third grade is the first year they get tested, so it has just shown up.”

“But,” I persisted, “you trust the people you hired to run the school. You spend a lot of time there too. How come they never noticed that the kids couldn’t read?”

“We all somehow deceived ourselves, I guess,” she concluded.

I proposed that we visit the school together and find out for ourselves if the students could read. We would rely only on the direct data: we would listen to a sample of students actually reading.

And we did just that. We randomly picked nearly one-fifth of the students in first through fourth grades and read with them. We did not find a single nonreader. Most of the children read aloud from “grade-level” books they had never seen, and “with expression.” They made reasonable efforts to pronounce unknown words and car-
ried on good discussions about the text. And they also knew a lot about books, authors, words, and language, although they were probably not yet as proficient as their more advantaged peers.

A week later, still stunned by these results, my friend said to me, “Debby, could we have been wrong?”

Well, maybe our sample was skewed, I suggested. No, she insisted, she had checked on that. Still, somehow . . .

We both laughed.

What she had on her hands was a political, not a reading, problem. The school’s future funding required better scores. She needed techniques that would raise their scores while not interfering with the reading performances that we had witnessed. The challenge was to figure out a way to make sure the children’s current low scores did not undermine their self-esteem and their continued impressive progress as readers.

It should not have surprised me when a year later my friend called to tell me that we had been right. The proof: “Our new test scores just came out, and they confirm our judgment.” But even she needed this additional “evidence,” although she knew that what had changed was not the children’s ability to read but their ability to demonstrate it on this particular test. Test scores still seemed more “real.”

WHAT STANDARDIZED TESTS MEASURE

If my personal account is not aberrant and my friend is not an oddball, then there clearly is a problem. This nation is spending inordinate amounts of energy trying to produce not better and more enthusiastic readers, but better test takers. No independent evidence exists that the two are the same thing or are produced by the same methods. The gradual replacement in schools of all other forms of judgment by this one particular mode of assessment, despite decades of attack, cannot be merely a misunderstanding or a shortcut born of economy—although both are essential to the deception. If the existing model were viewed as not merely harmless but dangerous, a more expensive solution might be found. After all, beloved medical cures have been discarded, though not easily, once they have been acknowledged to be bad for our health.

Yet a twenty-year effort to replace the traditional standardized test with more authentic forms of assessment—exhibitions, portfolios, oral defenses, performances, and the like—has foundered not merely on the grounds of expense but on the seemingly irreconcilable conflict between what is authentic and what psychometrics will allow. The conventional view is that only psychometrics can speak with objective certainty and that all tests must meet their criteria. As we will see, this view has made direct assessment appear almost impossible.
The Role of Rank Ordering

The stubborn fact may be that the particular rank order that psychometrically reliable testing alone seems to produce remains both politically and educationally useful. Our society tolerates such testing not because it is scientific, but because many people want a rank order—that is, some form of normal, or bell, curve. It seems to confirm a commonsense observation that since most of those in positions of power and influence score well, something important is being tested. Furthermore, there are just enough exceptions—surprises—to make those with social and political power feel that such tests cannot be entirely self-serving.

But above all, perhaps traditional testing allows researchers, educators, and policymakers to do things that otherwise would be very hard to do. This relatively cheap instrument makes it possible to diagnose and compare with elaborate tables and charts children, schools, and systems with fairly predictable results (Meier, 1974). By conforming to a set of “industry” rules, the test makers claim the right to promise the public that the score of a test taker would not change much if the person took a similarly constructed test, or took the same one on a different day, and, furthermore, that the results can predict this test taker’s success at similar kinds of tests taken at some time in the future.

Although the quality and quantity of the procedures that must be met are certainly greater for psychometrically well-designed tests than for other forms of judgment, the mystique of their infallibility far outstrips what the actual design can support. There is clear evidence, for example, that the results across different, equally “valid” commercial tests do not match each other, even though the rules of psychometrics would make this seem impossible. Furthermore, measurement error is far too great for the high-stakes decisions it is used for. Worse still, when children are read the questions aloud, removing any need to know how to read, one would suppose that their scores would go up substantially. The tests may not even be measuring the skill in question. But for many low scorers, this makes little or no difference. The test is often not measuring “knowing how to read,” but rather the children’s difficulty comprehending certain kinds of subject matter and vocabulary and their different interpretations based on different experiences. The “distracters” (wrong answers) are, in fact, intended to serve that latter purpose: to pick up on common misinterpretations that children might reasonably make. Thus the irony. Machine-scored multiple-choice tests can be 100 percent reliable even though they may be measuring something different from what was advertised.

Impact of Coaching to the Test

In the “olden days”—when I went to school and when I began teaching—it was the tests that schools and teachers designed themselves that were considered the centerpiece of teaching and learning and that formed the basis of in-school
authority. When external, norm-referenced tests were administered, schools and teachers were strictly prohibited from preparing students for them, except at the beginning of the actual test session and then only by repeating the explicit script the test maker prepared for this purpose. Students were told there was no way to prepare—except with a good night’s sleep. On testing day, teachers instructed students not to guess what the right answer was, even when, in fact, wild guessing could substantially improve one’s score and there was no penalty for wrong answers. As I recall, we teachers believed that the scoring machines could detect guesses, just as it is now claimed the tests can detect cheating (too many erasures, perhaps, or when the erasures always result in a change from wrong to right answers). These alien tests, like the equally secret world of IQ testing, belonged to a different domain—the domain of “science.” Even the scores were kept secret from the general public (although parents were informed about their own children’s scores).

When I began teaching I obeyed all these instructions, albeit with increasing reluctance, because I was told that only thus could the playing field be made level for all children. During the 1960s, however, test scores became public. Newspapers began to report scores, comparing schools and systems to each other. The race was on. Eventually parents, teachers, schools, and school systems began to realize they could improve students’ scores with better and longer preparation. Leading the way were for-profit organizations (such as Kaplan and Princeton Review) that tapped into the anxieties of more powerful parents over the emerging influence of the Scholastic Aptitude Test (SAT) in deciding who got into which prestigious college. As public pressure on schools to account for results grew—even for traditionally low-scoring students—the users of the regular standardized K–12 tests stopped obeying the publisher’s prohibitions, at first covertly and eventually by systemwide decree. By the 1970s, what is now benignly called curriculum alignment began to occur. Schools went into the business of designing what and how they taught in order to match the tests. For example, teaching the right answers to the items likely to be found on a reading test became the way to teach reading. Correlation became causation without missing a beat.

In the mid-1960s when I arrived in New York City, most city schools showed that the vast majority of students were scoring below the fiftieth percentile in reading and math on the MAT (Metropolitan Achievement Test). The city press had just begun its relentless annual public ranking of schools based on this test. The schools responded in due course so that within a half-dozen years the number above grade level rose dramatically. At the same time, no one claimed that public school children in New York City had made a breakthrough in reading, and no high school reported that it was getting better-prepared students. Districts unabashedly offered special programs aimed specifically at students performing just below “grade level.” (Interestingly, the scores nearly always plummeted when students made the transition to middle or junior high school.
The absence of a homeroom teacher accountable for both teaching and testing reading may be one reason for this otherwise inexplicable drop.)

Given the nature of the instrument, however, teaching to the test can only temporarily improve the odds, since only rank order counts, not one’s absolute score. For example, if Mary gets seventy-two answers right on a fifth-grade test, she falls into the sixty-third percentile, which is her reported score. (Other forms of reporting data may be included, the most common being “grade level,” the least common being stanines, which is a more accurate portrayal of the meaning of scores on the bell curve.) The reported score presumes that this is where Mary would stand if the entire population of fifth graders had taken the test. But by this logic it follows that improved scores based on improved preparation eventually requires the test maker to renorm the test, by changing either the score sheet or the actual test.

Even with renorming threatening the future, no school or school system could afford to ignore the temporary advantage of test-specific preparation in a society that judged its schools by their students’ test scores. Superintendents, after all, rarely lasted long enough in one position to worry about renorming. In contrast to older assumptions, it was now implicitly acknowledged that a student taking two different tests, each of which claimed to test reading, might not get the same score on both. And, in fact, it became general knowledge within the field that certain tests were “harder” than others for certain populations. Furthermore, it turns out that it makes a difference if one coaches to a particular test. Coaching “in general” is far less useful than coaching to a particular test. Finally, it makes a difference when a test was last normed. The reason that testing programs are sometimes switched by school systems is worth examining. It often coincides with a new school or school administration that has a short-term stake in demonstrating that students are not doing as well as the public may have thought and that the new team thus faces enormous difficulties based on past poor leadership. This is especially the case where new administrators honestly face what they know are inflated test scores based on past overpreparation for a particular testing program.

Efforts to raise scores at all costs would at least be harmless if the tests themselves were a good way both to teach and measure reading. But the nature of both the tests and the coaching misleads us and misdirects teaching. The trouble is that the low scores misled educators into spending more time and resources teaching the basic skills of reading and in coaching for the tests and less time on rich subject matter (history or science) or good literature. Concern for test scores led New York City in the 1970s to eliminate elementary school libraries and replace them with high-tech “reading labs.” Time is finite, so we spent our time on what literally counted.

Reading is often only a small part of the children’s problem. When a few researchers many years ago asked children why they selected a particular
answer, their responses suggested that something quite different was going on. That “something” may be closer to the mysterious quality that IQ tests also tap, which my ancestors, among others, appeared to lack three generations ago when they first came to America in droves. Jews and southern Europeans scored near the bottom on early IQ tests at the turn of the century and quite a few years into the twentieth, while northern Europeans scored near the top. Asians scored very low too, despite the current view of their talents. But years of immersion and success in American life erased those disparities.

What these researchers discovered is the same thing that I learned in the late 1960s when I tried to find out why my bookworm son scored so poorly in his third-grade reading test in New York City—a discovery that was confirmed in my interviews with children in Central Harlem, which I later wrote about (Meier, 1974). Even if they could read the original item aloud accurately, summarize it orally, and interpret it sensibly (as we had asked the children to do in my friend’s school), they might not get the right answer on the test. Poor children and minority children in particular gave what appeared at first glance to be irritatingly “wrong” answers. What the test missed is what can be learned only by questioning the students. Although the children were reading quite well, in the usual sense of that word, they were interpreting the text differently, or were baffled by particular references, vocabulary, and subject matter assumptions. These factors affect some subgroups of children far more than others and can have a substantial impact on a child’s test scores on any given test day.

In short, test scores are not merely weak indicators of higher-order skills, or affective and aesthetic strengths, but even of what we ordinarily think of as just plain “reading comprehension” skills. The kinds of items that would pick up gross errors in literal reading—the straightforward “know-how” questions—are intentionally few since a test that rests on such gross distinctions would not produce a normal curve. Thus the vast majority of what schools successfully teach and children successfully learn is intentionally left off such exams.

It was this form of testing—designed much like IQ tests—toward which the original prohibition against coaching was directed. The test makers needed items that would allow them to predict ahead of time how many children would get each answer right and wrong; they needed items they expected to stand the test of time (for at least a decade or more). The tests would work only if one had reason to believe this claim. The accuracy of the test makers’ predictions depended on items that schools were unlikely or unable to teach directly, but which differentiated children from each other. That is precisely what enabled these tests to be so useful: to predict future performance. Even to try to coach children on items intended to serve such a purpose would contaminate the scoring system itself and make all scores unreliable predictors of the future. This explains why the SAT was originally not called an achievement test but an “aptitude” test. Coaching therefore undermines precisely what the test makers have
so carefully designed. (Consider my amazement when, teaching in Project Head Start in 1965–1966, I was told that our goal was to raise IQ scores and that we should therefore focus our curriculum on teaching the specific kind of knowledge that was used on IQ tests. It went against everything I understood at this point in my life about the meaning of IQ.)

The Fallacy Revealed
What is puzzling is how it was ever imagined that such an instrument could measure the nation’s progress at creating a better-educated citizenry. What psychometrics was offering was a promise of “no change” over time—just a picture of who belonged where along the immutable bell curve. Americans were, to their credit, unwilling to admit that a child’s socioeconomic status was already a better and cheaper predictor of a child’s and school’s place on the normal curve. In the society of the past, where school success was only one among many ways to succeed, the issue of exactly what was being measured could be bypassed without too much fuss. No one seemed wildly distressed by the secrecy surrounding testing until the 1960s. It took the increasing importance placed on school success and the 1960s thrust toward social and racial egalitarianism to force the issue into the open. Even so, few dared to claim what virtually all educators and the larger public had come to realize: that schooling was powerless to eliminate the advantage—the edge—that comes with being advantaged (never mind uglier racist or ethnic assumptions about where groups would stand on the normal curve). Indeed, as the producers of Sesame Street discovered in their early assessments of the program’s impact, the already advantaged are more, not less, likely to take advantage of the best and newest learning schemes. Even fewer people were willing to acknowledge that leveling the playing field would require far more radical and targeted reforms than those being proposed (and not just changes in schools), beginning with abandonment of a form of measurement that promised no change. If a way were discovered to get all children to read better, faster, and earlier, nothing would appear to have changed as long as children were measured by these kinds of tests. Such a breakthrough would produce a difference only if the frontrunners were refused access to these same new and better learning schemes. That is hardly likely to happen, nor should it.

When Coaching Is an Abuse and When It Is Not
If the public wants tests that produce a reliable and predictable rank order, then it has them, and people should stop abusing and misusing them with coaching.

Coaching to change scores, even if it often succeeds in this respect, far less often improves reading and can even endanger good reading. For one thing, it
masks the real problems that low scorers are having, which more often lie beyond “reading programs” and the whole-phonics-versus-whole-language debate. On the surface a test can be about the skill being tested, but the connection between that skill and what students are having difficulty with can be very different.

Of course, coaching for tests is not always an abuse, depending on the test and its purpose; it can even be quite appropriate for certain kinds of tests. Coaching is standard for a driver’s test, because that test is not designed so that coaching will destroy its scoring system. There is no mandate, for example, that scores must follow a particular curve. A road test simply and literally measures driving competence: the more people who pass, the merrier. The preparation is in keeping with the nature of the task. In contrast, preparing for an eye test (by prior access to an eye chart) would not serve us well. The purpose of this test is strictly diagnostic, and coaching would cover over a problem that needs uncovering. Its purpose is not to determine how well Americans are “seeing” absolutely and comparatively; rather, it is to help each particular child see better. Wouldn’t it be nice if reading tests did the same? At least occasionally?

In standardized school tests, however, science has given us tests that produce a reliable, predictable rank order that does not change over time and that matches predicted future success at schooling—as long as the test is not abused.

What the Public Has Bought

Coaching on standardized tests is an abuse if the users—the audience—think they are getting something else. What the public thinks it has bought is a combination diagnostic instrument that can inform teaching and learning, a measure of whether an individual child can read and what percentage of all children can do so. They think that what the test makers are offering them is more or less like a driver’s test, separating those who can from those who cannot. Especially where their own child is concerned, they want the chance to check the results, to get as close as they can to what went “wrong.” But since the questions and answers on the actual test must remain secret—unlike our driver’s test—it is impossible, without cheating, to “show” parents, teachers, or the child himself or herself what the mistake was, meaning the nature of the ignorance in question. Yes, of course, parents and laypersons would also like a single unambiguous and simple number that stands in for all this complexity. But they want it only if it is possible. Simple information is what they got, not honest information.

Although it is easy to point to parents, lay citizens, and politicians as the culprits in their wholesale dependence on traditional tests, there is more to it than this. In a world where expertise is honored (how else can one survive?), the blame lies at least partly with the educational statisticians, researchers, and psychometrists who lent their names, reputations, and work to accrediting such data, even in their much-abused and misused forms. If these test data were good
enough for these experts and all their task forces, is it so surprising that Mom and Dad and the average school board member would take them seriously?

It was the professionals, above all disinterested ones like researchers and scholars, who offered the general public psychometrics nearly a century ago as the answer to all its desires: short, sweet, cheap, simple to rank, seemingly precise and certain, and with results that confirmed society's expectations. These professionals bear some responsibility for what the public now thinks it can have—for example, that the test score gap between high and low scores on standardized tests can be closed. Just as good doctors do not pretend to certainty, even on matters of life and death, educators should not have done so either.

Any test has its own parameters, its own rules, requiring one to ask particular and different questions. Out-of-classroom educational experts and researchers were and still are needed to act as guardians of the information pool on which wise decisions have to be made in a democratic society. Instead the experts have helped contaminate that pool. By both their actions and inactions, members of the profession have not only encouraged the growth of such tests, but have allowed standardized norm-referenced tests to serve a myriad of conflicting purposes that no single test, even a good one, can possibly fulfill. A test that was intended to serve individual parents, classroom teachers, the school district, politicians, and the lay public does not, it turns out, do well by any of them. Parents and teachers have mainly needed diagnostic information: How well are the students getting what it is they are being taught? In what different ways do different students learn best? That is not of much interest to politicians or the broader lay public. And none of these tests is designed well for sorting people on the basis of the particular skills they need to succeed in their institution, trade, or occupation, be it as a police officer, an undergraduate at Harvard, a student at medical school, or a member of a drivers' club. No single test can sort for all these high-stakes purposes, each of which rests on a different set of dispositions, habits, and skills.

In addition, the degree of measurement error in standardized tests is considerable; it is certainly too large to serve us well where the stakes are high and when decisions will be made that determine the future course of a young person's life. The younger the children are, the greater the measurement error is, yet we are seeing more and more testing of very young children. Colleges and employers may find they have institutional needs for arbitrary cut-offs that cannot be justified by statistical evidence, but educational practitioners do not have to use cut-offs for their own teaching and learning purposes—for decisions they make between kindergarten and twelfth grade. A 3.3 score and a 3.9 score are statistically the same. Yet in many localities a child with a 3.3 gets automatically held over, while a child with a 3.9 is patted on the back.

Test data, by the way, are not the only form of "hard data" that have misled the public about the performance of schools. New York City for years counted
graduation rates by calculating the percentage of twelfth graders who graduated. This meant that in places like the Bronx, where little more than a third of the students ever made it to the twelfth grade, the press would on occasion release Board of Education data claiming 90 to 97 percent graduation rates! Who knows how “dropouts” are figured? Some cities do not call students dropouts who leave after reaching the age of eighteen; others do not if the students leave before they reach high school. Many schools and school systems disguise dropouts by checking off that the students are moving to another state or country—Puerto Rico, Mexico—if at all possible without regard to whether they are likely to be continuing their schooling. I leave aside the many ways that principals can, if enough is at stake, play with attendance data. New York City once upped its citywide attendance records by simply counting noses at a different time of day. Imagine my surprise to discover that in California and Massachusetts, daily attendance data include as “present” students who are not in attendance if they have a note from home. National comparisons of “incidents” and suspension rates are based on wildly different criteria and frequently depend on particular local consequences; more incidences reported may lead to more security guards or more accusations of racism and bad management.

What the community of educators can and should be faulted for is having let this kind of stuff appear to be hard data just because some data are better than none.

TOWARD A NEW PARADIGM

Some of these faults—like how to count dropouts and attendance—could be fixed, although this would require the cooperation of a better-educated press. But the student testing model now used for charting progress is unfixable. Painfully needed is a new paradigm to replace norm-referenced testing. The one we are using is part and parcel of the historical fact that schools have been as much as anything else about sorting children. The capacity of scores to predict the future was built around the assumption that schools could not and should not have an impact on rank order. Now that we claim to want something different from our schools, the technology we invented for a different age is grossly unfit to do the job.

Our society is accustomed to tests that serve as scientifically neutral stand-ins for decisions that many other societies make solely on the basis of wealth or social status and others delegate to people they trust. Democratic ideals and a healthy antihierarchical bias were in part responsible for bringing into existence the granddaddy of all modern psychometrics, the American IQ test. But as Stephen Jay Gould has so dramatically demonstrated in his The Mismeasure of Man (1981), IQ tests came with their own baggage of biases. All of our most
reputable standardized tests, no matter what they purport to measure, have been modeled on the IQ test. There are more sophisticated ways to get the samples, score the tests, and report the results; there are even some new gimmicks. But there have been no breakthroughs in the basic technology of "scientific" testing in nearly a hundred years, and none is to come.

Rather than the technology, it is the idea on which the technology rests—the idea that there is a way around the fallibility of well-educated judgment—that is at fault. Standardized norm-referenced tests stand in stark contrast to what a democratic school system should be all about—what Thomas Jefferson called schools that educate our "discretion." They ignore the aspect that democratic society ultimately rests on: an educated citizenry trained to exercise judgment, with all its fallibility, uncertainty, and messiness, a messiness kept in check by a combination of healthy skepticism and public debate, not a pretense to certainty built around a network of secrecy.

At a time in history when it is claimed that our society must educate all its children well and that all children must be trained to use their intellects in sophisticated ways—a proposition contrary to the concept of normative tests—the continued popularity of this outmoded form of testing among those who know better needs addressing, and in a hurry. Some of the technology of psychometrics can be constructively used for testing small, random, uncoached samples of students in order to make comparisons as well as track changes over time. Such use, without high-stakes consequences, would not require norming; nor would measurement error matter. Studies could be made of the reasoning behind the ways that different groups of students respond to the same questions in order to learn something from such responses, not just mark them wrong. This kind of research could assist us in teaching, rather than labeling, children.

But beyond the more constructive use of such tests, there need to be different approaches to assessing students and their schools—approaches that speak directly to parents, teachers, and communities, helping them to understand their children better and track their progress over time. We have reached the point where the traditional testing system threatens all schools, particularly those that serve the most vulnerable.

As long as the man and woman on the street and powerful policymakers believe they can answer profound questions of student capacity, teacher efficacy, different forms of pedagogy, and even the impact of family structure (to mention just a few) by analyzing in myriad statistically sophisticated ways the latest available test data, we are in trouble. Instead of recognizing this, however, we are in the midst of a new and more far-reaching effort to use such data for even more grandiose purposes.

Currently policymakers are demanding nationwide improvement in test scores (or else); many are pushing for the creation of one uniform set of interconnected nationally norm-referenced tests to replace all the other forms of con-
fusing human judgments. They think that such an airtight system of testing will finally bring coherence to an overly decentralized education system, put pressure on localities and lazy practitioners, ensure that all children are measured fairly (and frequently), and make it hard for anyone to rely on their ill-or well-intentioned biases. In the process each child, family, and teacher will finally have the truth about where they stand in the grand scheme of things and what they have to do about it—of course, raise their score. We have finally squared the circle.

Such proposals can only help the exodus from public schooling altogether by those with the means to flee. Centralized school systems with their sure-fire consequences may be okay for other people’s children, but they are unlikely to appeal to users with choices.

Building the New Paradigm

Of course, there are alternatives and lots of good folks who care and believe in them. The problem is not that there are no better ways. It is actually not so hard to build assessments that “merely” help us see where students stand in a continuum or assortment of skills. In fact, they were invented decades ago, have been used for ages by school people, and are still around. Standardized criterion-referenced testing, which came into vogue in the 1970s, was intended to do something of the sort. Unfortunately, criterion-referenced tests could not resist psychometrics, and in the name of reliability they ended up as simply longer traditional tests with more subscores.

By the mid-1970s, some reformers went a step further. They restated their alternative ideas on other familiar forms of assessment, most of which divided learners into only two categories: pass-fail. They looked at how Ph.D. candidates are judged, how architects are certified, and at older forms of public exhibition—from the Boy Scouts, to 4-H, to bar mitzvahs. They noted that driver’s tests matter a lot, but no one gets ranked. Oddly enough, pass-fail is all that counts in the bar exam, although preparation for the bar is much joked about by those who have passed the exam for its irrelevance to lawyering. Red Cross lifesaving tests are pass-fail. My own children had to be able to swim halfway across the lake before they could take the canoe out on their own. They could do it with speed and form, or they could comfortably float and dog paddle. There were some uncertainties that required retesting, where my husband and I disagreed on how “comfortably” my son was dog paddling.

There are circumstances, school reformers of the 1980s noted, when “coaching” for tests makes sense. Ted Sizer, chairman of the Coalition for Essential Schools, and Grant Wiggins, a much-published author and expert on secondary school assessment, argued more than a dozen years ago for schools built around exit exhibitions of competence. They saw good coaching as precisely the task
of good teaching. There was nothing wrong with aligning ends and means. Good teaching always has this in mind.

An important national movement in the 1980s was built to try to weaken us away from the old paradigms to new ones built on authentic forms of performance. There were, and are, differences of opinion between the proponents of these new paradigms on many issues, but all have agreed that their work stands in stark contrast to the older psychometric model. Their work fundamentally rests on the concept that it is possible and desirable to review actual novice performances relying on the judgment of one’s peers or would-be peers (the experts). Many proponents of such forms of assessment would even argue that becoming adept both at exhibiting one’s competence in this fashion or at being able to make judgments about such exhibitions is at the heart of their definition of a good education.

Advocates of traditional “scientific” testing might suggest that these systems of making judgments are too “political” and sloppy for making important life decisions. But if the central purpose of schooling in a democratic society is to train fallible human judgment, it makes sense to assess schools in ways that do honor to this essential purpose also.

Maybe the obvious alternatives are avoided for precisely this reason, because they require us to describe what is good enough and then stand behind it, rather than rely on nameless others to set our standards. In the 1980s and early 1990s Vermont pioneered a statewide assessment system that was built around collections of student work. The exciting thing about this short-lived portfolio approach was that it required all the state’s teachers to look at the work youngsters produced in the classrooms of their colleagues and make judgments about it. It required, in other words, that teachers in the state take responsibility for their opinions. Such approaches to assessment hold us all accountable in deeper ways. They provoke us to wonder whether most of us are “good enough” and, if not, whether it matters. They stimulate debate about the purpose of public schooling, a debate that would be both hard work and inevitably contentious and might not lead to consensus.

So we pretend that it is all pretty simple. After all, doesn’t everyone want a school that “works”? Is that really so hard to define? Isn’t that where most children get high scores? Aren’t the scores indicators of doing well at school, and if they are not, can’t school be reformed so there is better alignment? Soon enough, scores define purpose. In the absence of better answers, why not?

When asked why one must go to high school, most young people will answer, “To get a diploma so I can go to college.” Why go to college? “So I can get a better-paying job.” Rare is the student who says of either high school or college, “So I can learn something of intellectual or social value to me.” The value of the certificate lies not, therefore, in what one has learned but in its
scarcity. Its value can be raised by making it harder to pass—by devising either a tougher test or a tougher scoring system. High-prestige schools have higher cut-offs (pass-fail dividing line), low-prestige schools low ones. No one ever has to remember or agree on what it is we value so highly that we require all our children to devote nearly two involuntary decades to mastering it. The test is a hurdle, not an accomplishment.

“All Children Can Learn”

A new idea has been unleashed: all children can learn things of intellectual importance. It is still an amazing claim, and in the world of tomorrow perhaps a critical one. It is surely the most exciting national challenge our country faces. But it will be a cruel hoax if we do not alter the way we measure all our children.

Tackling purposes can change this potential hoax into a realizable dream. That is why it has been at the center of the best reform ideas. It was the central point of Ted Sizer’s much-acclaimed Horace’s Compromise (1985), Neil Postman’s The Ends of Education (1996), John Goodlad’s A Place Called School (1984), and many other books that have encouraged a healthy, though not widespread, debate about the essential purposes of schools. From 1975 to 1995 governors, think tanks, national task forces, and disciplinary leaders all began to echo at least some of the rhetoric of these writers. Among the many initiatives in the early 1990s was a widely acclaimed national effort to spearhead systemwide reform launched by President Bush and U.S. Undersecretary of Education David Kearns (former chief executive officer of Xerox Corporation). Called the New American School Development Corporation (NASDC), it invited reformers throughout the country to compete for the prestigious designation as a NASDC model. There were remarkable similarities among most of the models proposed, as well as among the final five selected for full funding.

Whether launched by professed liberals or conservatives, phrases like higher-order thinking or critical thinking, problem solving, teamwork, learning for understanding, and performance-based assessment, along with decency, character, compassion, and awareness of the ideals and feelings of others (the diversity stuff) were increasingly “in.” People who disagreed on many fronts joined forces in praising the National Council of Teachers of Mathematics (NCTM) frameworks, instruction in reading that rested on the use of good literature, and a science education curriculum that rested less on rote memory and more on learning to experiment and reason scientifically. No one, except what was seen as the extreme “Christian right,” blinked for a while. The dissension came over pedagogical details: the balance between memory and discovery, the role of group versus individualized teaching, more phonics or less, and what constitutes evidence of success.
During this same period, cities and states as well as a variety of interesting national reform networks initiated changes in the way students were being educated and schools organized. Unlike overhauling a factory, changing schools had to take place amid the arduous daily tasks confronting schools and teachers. We could close a factory for retooling but not a school. School change also bumped up against ordinary people's ideas of what made sense, their own memories of past experiences at school, their fears about novelties and experiments being practiced on their children. And, finally, school change had to respond to the sixty-four-thousand-dollar question: "How do you know this new way is better?" All of these factors had to be answered before changes were even launched, much less fully established. It is a problematic task at best. But it did not help when we agreed, however begrudgingly, to answer the last question—"Does it work?"—with instruments that came out of a different history to serve a different purpose.

To the latter question the experts and school practitioners who were seeking serious change had some alternate, if hard-to-sell, answers. Public exhibitions, performance assessments, portfolios, school quality reviews by external experts, and longitudinal data based on real-life success were suggested. They were well accepted in many other fields of work, and with enough effort from the highest authorities and sufficient time, the public might have bought them as good ways to assess their schools. A rash of states and publishers began to play with how to do this on a large scale, while on a smaller scale individual schools and teachers began to try out their own forms of defining and exhibiting achievement. California tried on a large scale to build a new assessment system that was standardized in some traditional ways but deviated radically in others. Vermont went even further and introduced statewide portfolio assessment. Kentucky tried a statewide system of authentic testing. New York State launched a task force to develop a new approach to statewide regents' testing that would combine some form of standardized tests and some locally devised alternative assessments, along with schoolwide peer review that borrowed from British practice. Even the testing of teachers launched by the National Board for Professional Teaching Standards opted for portfolios, video observations, interviews, performance reviews, and written essays rather than standardized multiple-choice testing, and it refused to consider scoring by the bell curve, despite the enormously increased cost and time involved. Highly reputable folks in the various academic associations agreed to develop similarly sophisticated forms of student assessment in each discipline.

All claimed to be in agreement that the old tests—whatever their virtues—could not tackle the lofty new missions. What was needed, if the nation were serious about wanting the best for all children and their teachers, were new but rigorous ways to assess student and system progress toward meeting these goals.
The new formats exposed the enormous weak underbelly of America's schools—how little of what we taught was exhibitable. In schools in which students were rarely writing and rarely talking (at least about schoolwork), what was there to exhibit? The reformers viewed this as one of the positives—that such new forms of assessment might drive schools into rethinking their practices. Children tested on performance would have to be taught how to perform.

This in turn led to an interesting nascent debate about what kind of ultimate performances students were to be prepared for. Who were the audiences? The connection between tests and performances on real-life tasks by people outside of school settings became more relevant. The scores of traditional tests, in their grand abstractedness, masked such connections and left fuzzy the relevance of in-school training to out-of-school purposes.

Subverting the Idea

In the midst of this very fragile and healthy debate about purposes, and while the lay public remained largely ignorant of what the fuss was all about, the test makers and their allies noticed a flaw: these new forms of assessment, which depended on the educated judgments of thousands of practitioners, including sometimes even parents and lay citizens, did not mesh with the world of psychometrics, and any attempt to make them do so would be inordinately expensive, not least because they would require the retraining of teachers and test givers. As constituted, they lacked a scientific base. In fact, the more authentically the trait assessed was, the more it involved the people closest to the children—sometimes even the children themselves—the less it met such a requirement. In "real life" these judges were too much like other jurors—like movie critics and publishers, often disagreeing, and sometimes in substantial ways. In short, the new forms of assessment were more expensive, they were harder to explain to the public, and they lacked what the old tests had: instant familiarity. The policymakers responded to these arguments by demanding that the nascent experiments conform to more standardized formats: less choice of topics, clearer guidelines for essays or scientific experiments, more standardized questions (that were also less authentic). They agreed that "better" rubrics with less wiggle room should be designed and that scorers needed to be trained to produce more consistency—to act more like the machines they were replacing.

In addition, as the late 1990s rolled around, attention began to focus on whether the cut-offs were tough enough. By this time, "tough" had become synonymous with cracking down on precisely those closest to the students—their permissive or misguided or ignorant families and teachers. The steam behind the idea of building a new paradigm was gone. The old tests emerged stronger than ever. (An odd test that has all the faults of both paradigms was also born of this new mood. Like the traditional tests, it is entirely paper-and-pencil, are
very long, and claim to be able to arrive at highly precise numerical scores with indisputable cut-offs. They are based, however, not on years of careful reliability or validity studies, but on the judgments of appointed and centralized panels. I speak here, for example, of the new Massachusetts student and teacher tests that seem to be leading a new trend in testing, which needs attention from the research community for its equally serious potential for abuse and misuse.

It is hardly surprising, then, that the sentence “All children can learn” has finally been filled in by the ludicrous “All children can learn to pass any sound psychometrically designed test with equally high scores.” Lake Wobegon has gone national; and, if not, someone is just not working hard or smart enough.

The fragile reform consensus began to fall apart. Literature-based instruction, nontracked classrooms, interdisciplinary instruction, and even the seemingly unassailable NCTM standards are increasingly under bellicose attack for failing to produce improved test scores. The attack comes not only from places like the conservative American Enterprise Institute but also from the mainstream American Federation of Teachers. Nervous supporters of these past reforms are responding by watering down the boldness of their ideas. Overnight, school people are being given new messages in direct conflict with the ones they were scolded for not embracing last year (probably most breathe a sigh of relief). Restructuring buildings, not changing the way teachers and children work together, is now “in.” And cynical teachers know that this too shall pass.

As long as tests are designed to produce a set rank order, children and schools can at best change places with each other. If the demand is for “tougher” standards, schools can raise the cut-off point so that fewer pass. But meanwhile the dumbing down involved in the organization of schooling around getting the “right” answers on such tests will continue to be ignored. Over time a new consensus may develop that concludes that it is the mantra about “all children” that is passé, not the tests. The “realists,” with an eye on the political desirability of cutting expenditures, will be able to claim that the bell curve is destiny.

**RETHINKING THE GOALS**

What could change this gloomy prognosis? If we wanted all our friends and neighbors, nay even our enemies and potential rivals, to be “equally” well educated, and if we believed that there might be more than one “politically correct” definition of well educated, maybe we would be willing to give up on our current dependence on psychometrics and all that goes with it. Maybe then we would turn to some of the not-so-complex alternatives that already exist: both old-fashioned school assessments like exhibitions and the myriad assessments used in other fields of endeavor. Maybe we would recognize that the heart of
any good assessment system is that it fosters the kind of healthy intellectual debate that democracy depends on. In this context, psychometrics might have a contribution to make, but it would not dictate the terms of the debate. In such a context, school restructuring would not be an end in itself but simply a tool for creating the kind of lively exercise of judgment needed to raise children well. Big schools versus small ones; interdisciplinary versus disciplinary; teamwork versus solitary work: all would be debated in relationship to the purposes of education, not their impact on test scores.

But unless the emperor is exposed, why should the public invest in an expensive and more complex alternative? If one seeks to adopt the novel and the exciting idea that all children could and should receive an education as demanding and as powerful as the one largely reserved for a small elite, this will not be easy to realize. It is a fragile, still untested, and hardly well-funded idea, and it will remain unattainable by statistical fiat if current assessment practices are not rethought, if the “stuff” out there away from the streetlight does not become the center of our attention—the stuff that will get us talking, even arguing, about what it really is we want our children to become. What we need is a frank and occasionally raucous debate about the habits of heart and mind that we hold ourselves publicly accountable for nurturing in the next generation of our fellow citizens, within a spirit of compromise and tolerance for more than one answer.

The research community will need to become part of this debate. The argument I make in this chapter is that researchers have been using existing “hard data” as the base data of much of their research on educational practices and student performance—this despite the fact that such data are at best corrupt and at worst built around a false paradigm, false at least with respect to the purposes for which it has been used. Not only does this make much of their research questionable, but it has played a major role in making the test scores respectable.

At a time when the overuse of testing is drowning out good practice and when testing is actually cutting off debate about purposes, researchers could provide a powerful voice of sanity. Good conversation about the connections among purposes, practices, and assessment will not lead to simple answers—and it will be hard to conduct. But it might lead to a better-educated citizenry, on which better schooling ultimately rests.

References


