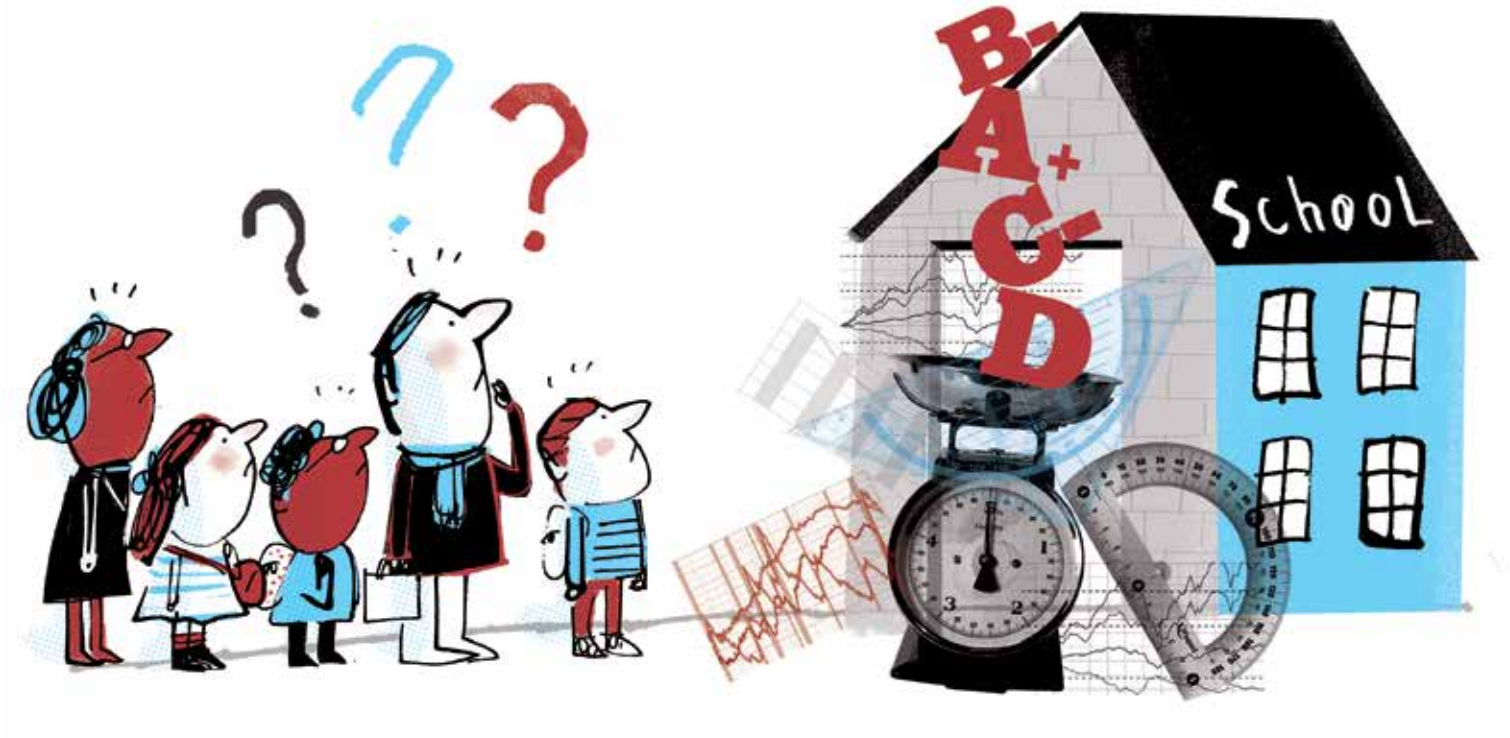
Moving Beyond the Failure of Test-Based Accountability



By Daniel Koretz

Pressure to raise scores on achievement tests dominates American education today. It shapes what is taught and how it is taught. It influences the problems students are given in math class (often questions from earlier tests), the materials they are given to read, the essays and other work they are required to produce, and often the manner in which teachers grade this work. It can determine which educators are rewarded, punished, and even fired. In many cases, it determines which students are promoted or graduate.

Daniel Koretz is the Henry Lee Shattuck Professor of Education at Harvard University's Graduate School of Education, a member of the National Academy of Education, and the author of Measuring Up: What Educational Testing Really Tells Us. *This article is an excerpt from his new book,* The Testing Charade: Pretending to Make Schools Better, *by Daniel Koretz, published by the University of Chicago Press.*

This is the result of decades of "education reforms" that progressively expanded the amount of externally imposed testing and ratcheted up the pressure to raise scores. Although some people mistakenly identify these test-based reforms with the federal No Child Left Behind Act (NCLB) enacted in 2002, they began years earlier, and they will continue under the somewhat less draconian Every Student Succeeds Act (ESSA) that replaced NCLB in 2015.

Examples abound of how extreme—often simply absurd—this focus on testing has become. In 2012, two California high schools in the Anaheim Union High School District issued ID cards and day planners to students that were color-coded based on the students' performance on the previous year's standardized tests: platinum for those who scored at the "advanced" level, gold for those who scored "proficient," and white for everyone else. Students with premium ID cards were allowed to use a shorter lunch line and received discounts on entry to football games and other school activities.[1]

Newspapers are replete with reports of students who are so stressed by testing that they become ill during testing or refuse to

ILLUSTRATIONS BY FRÉDÉRIC BENAGLIA

come to school. In 2013, for example, eight New York school principals jointly sent a letter to parents that included this: "We know that many children cried during or after testing, and others vomited or lost control of their bowels or bladders. Others simply gave up. One teacher reported that a student kept banging his head on the desk, and wrote, 'This is too hard,' and 'I can't do this,' throughout his test booklet."[2]

In many schools, it is not just testing itself that stresses students; they are also stressed by the unrelenting focus on scores and on their degree of preparation for the end-of-year accountability tests. Test-based accountability has become an end in itself in American education, unmoored from clear thinking about what should be measured, how it should be measured, or how testing can fit into a rational plan for evaluating and improving our schools.

The rationale for these policies is deceptively simple. American schools are not performing as well as we would like. They do not fare well in international comparisons, and there are appalling inequities across schools and districts in both opportunities for students and student performance. These problems have been amply documented. The prescription that has been imposed on educators and children in response is seductively simple: measure student performance using standardized tests and use those measurements to create incentives for higher performance. If we reward people for producing what we want, the logic goes, they will produce more of it. Schools will get better, and students will learn more.

However, this reasoning isn't just simple, it's simplistic—and the evidence is overwhelming that this approach has failed.

Ironically, our heavy-handed use of tests for accountability has also undermined precisely the function that testing is best designed to serve: providing trustworthy information about student achievement. It has led to "score inflation"—that is, increases in scores much higher than the actual improvements in achievement they are supposedly measuring. The result is illusions of progress; student performance appears to be improving far more than it really is. This cheats parents, students, and the public at large, who are being given a steady stream of seriously misleading good news.

Perhaps even worse, these bogus score gains are more severe in some schools than in others. The purpose of test-based accountability is to reward effective practice and encourage improvements. However, because score inflation varies from school to school and system to system, the wrong schools and programs are sometimes rewarded or punished, and the wrong practices may be touted as successful and emulated. And an increasing amount of evidence suggests that, on average, schools that serve disadvantaged students engage in more test preparation and therefore inflate scores more, creating an illusion that the gap in achievement between disadvantaged and advantaged children is shrinking more than it is.[3] This is another irony, as one of the primary justifications for the current test-based accountability programs has been to improve equity.

In *The Testing Charade: Pretending to Make Schools Better*, my new book from which this article is drawn, I document the failures of test-based accountability and describe some of the most egregious misuses and outright abuses of testing, along with some of the most serious negative effects. Neither good intentions nor the value of well-used tests justifies continuing to ignore the absurdities and failures of the current system and the real harms it is causing. My book, however, is not an argument against accountability. My experience as a public school teacher, my years as a parent of children in public schools, and my decades of work as a researcher in education have made clear to me the need for more rigorous and effective accountability in public education. But there are more sensible ways to go about this than the ones we have used in recent years.

Our heavy-handed use of tests for accountability has undermined precisely the function that testing is best designed to serve: providing trustworthy information about student achievement.

Here I present some options for doing better. In making specific suggestions, I frequently refer to "accountability."

I don't mean by this a system—like our current one—in which each school and often each teacher has one or more numerical targets and reaps punishments or rewards on that basis. Rather, I am using the term in the more general sense of monitoring how well teachers and schools perform and using a variety of methods to induce—and enable—poor performers to do better.

We Must Measure What Matters

The starting point has to be deciding what matters most. There is room to argue about this, and the list could become quite long, but I'll start with what I'll call the Big Three: (1) student achievement, (2) educators' practices, and (3) classroom climate.

There isn't much controversy these days about the Big Three. Even within the constraints of test-based accountability, many states and districts are trying out

ways of measuring both practice and classroom climate. There is, however, argument about how to measure the Big Three and about how much weight each should be given. In most districts, test scores still swamp everything else. Indeed, ESSA *requires* that test scores swamp everything else.

Let's start with student achievement. Perhaps surprisingly, given the many pages I devote in my book to all the flaws and unintended consequences brought about by testing, I'll begin by saying that standardized tests should be a part of any system of monitoring and accountability. Many critics of our current system blame standardized tests, but for all the damage that test-based accountability has caused, the problem has not been testing itself but rather the rampant misuses of testing.

The strongest argument for using tests in a system of monitoring is precisely the fact that they are standardized: ideally, students everywhere confront the same tasks, administered and scored the same way. This stands in stark contrast, for example, to high school grades, which vary in rigor from one school to another and even from one classroom to another. Standardized test scores mean—or ideally they *can* mean—the same thing regardless of where students attend school, and that in turn allows us to answer critically important questions, such as whether the achievement gaps between minority and nonminority students have really narrowed in recent years.

The rub, of course, is the caveat "ideally they can." The pressure of accountability has undercut precisely this advantage of standardized tests. Even leaving aside cheating, some schools engage in far more bad test prep than others, often causing comparisons based on scores to be completely misleading. For example, in some places standardized tests have created an illusion that the achievement gap between disadvantaged and advantaged students has narrowed far more than it actually did. That's because of high stakes, not flaws in the tests.

So, I should be more precise: we ought to start with standardized tests *if and only if we take steps to dramatically reduce bad test prep and inflated scores*.

What's the solution? Precisely what the designers of standardized tests have been telling us to do for more than half a century, and what the Finnish, Dutch, and Singaporean systems do routinely: use local measures of student achievement—that is, measures not imposed from afar. These local measures include both the quality of students' work and their performance on tests designed by educators in their schools, both of which go into the grades that teachers assign. In addition to providing a far more complete view of students' learning, using these local measures—along with standardized tests when we have good ones—would give teachers more of an incentive to focus on the quality of assignments and schoolwork rather than just preparing students for a single end-of-year test.

Beyond the Big Three, I'll add one more: what are often now called "soft" or "noncognitive" skills—attributes such as persistence, the ability to work well in groups, and so on. E. F. Lindquist, the same pioneer of achievement testing who warned that tests must be used in conjunction with local measures of learning, also cautioned—more than half a century ago—that skills of this sort that can't be captured by standardized tests are a critically important goal of education.[4] This may strike some hardheaded advocates of accountability as "soft," but recent research has begun to confirm the wisdom of Lindquist's advice: soft skills affect how well students do long term, even after they leave school. And research suggests that teachers' influence on these soft skills is distinct from their impact on students' scores. (For more on social and emotional development, see the article on page 16.)

For example, a 2016 study by Kirabo Jackson, an economist at Northwestern University, showed that teachers vary in their impact on absences, suspensions, high school completion, and later college enrollment, separate from their influence on test scores.[5] While it is not at all clear yet how measures of these outcomes can be incorporated into an accountability system, it is certain that we want to encourage teachers to help students develop them, and holding teachers accountable for scores won't accomplish this.

We Must Build a Sensible Accountability System

Measuring a broad range of important things is an essential first step, but it's not in itself enough to create reasonable incentives. I'll suggest four additional steps.

The first may seem self-evident, but it is routinely ignored regardless: *the system has to emphasize what's important*. The weight we give to various measures should, as much as possible, reflect their actual importance. It simply won't suffice to tell districts that they need to throw in one or more measures in addition to test scores. Unless the others are made to matter, test scores will still trump all the others. If the quality of instruction and classroom climate are truly important, educators need to know that they really count.

The second step is to *create the counterbalancing incentives* that are largely lacking in our test-based accountability systems. In our test-based accountability system, everyone, from a teacher's aide to the district and state superintendents, has the same incentive: to raise test scores. No one has a strong incentive to worry about *how* scores are raised—for example, to tamp down bad test prep. This is why districts sometimes provide bad test-prep materials and why administrators pressure teachers to use them.

The third step requires *looking well beyond what happens on any single day in the classroom*. William H. Schmidt at Michigan State University, who has devoted much of his career to international comparisons* of both student achievement and curricula, argues that in many countries, evaluations of schooling include monitoring how well educators are teaching the intended curriculum—that is, the curriculum that is supposed to be taught. One of the most common inappropriate responses to test-based accountability has been to stop teaching the entire intended curriculum, cutting back on or completely dropping whatever happens not to be on the test. The test essentially *replaces* the intended curriculum. To tamp this down, one has to compare what is called the "implemented curriculum"—that is, the content that is actually taught in a school—with the intended. This means that, from time to time, someone would have to examine teachers' syllabi, and often some of their lesson plans.

Monitoring how well the curriculum is taught is essential for a second, perhaps even more important, reason: it is one way to combat the impoverishment of instruction in untested subjects that test-based accountability has caused. A common response by educators to testing in a limited number of subjects has been to take time away from other subjects, sometimes virtually or entirely eliminating them from instruction. In the current system, no one has any incentive to tell teachers that a week of social studies isn't enough or that art class shouldn't be used to drill kids with math test-prep materials.

Finally, *targets have to be reasonable*: the goals facing educators have to be ones that they can reach by legitimate means. This requires practical targets for both the amount of improvement and the time allowed to accomplish it. The time span must take into account the year-to-year fluctuations in scores that arise from both differences among cohorts of students and the often unavoidable trial and error in improving instruction, because ignoring these makes annual targets a recipe for failure.

There is room to argue about how best to determine what is reasonable, but the principle is inescapable. If we demand more than educators can deliver by teaching better, they will have to choose between failing and cutting corners—or worse, simply cheating. This may sound obvious as a general principle, but, in practice, it will be both controversial and difficult to implement. Demanding big and rapid gains makes for good press and often good politics, so persuading policymakers to be realistic won't always be easy.

Use Tests Sensibly

Time after time, as bad news about test-based accountability began to accumulate, its advocates insisted that if we just substituted better tests—what they considered "better" varied from one instance to another—the system would right itself. They maintained that the negative effects on instruction and score inflation would be brought under control and that we would finally get the promised improvements in learning. This didn't happen, and while I don't want to disparage efforts to improve tests, these arguments missed the main story. The chief problem was never the tests themselves. It was the misuse of tests, which was often worsened by successive reforms.

We shouldn't rely on tests when we don't have appropriate and sufficiently high-quality tests to use. As much as is practical, we need to avoid relying on arbitrary performance standards, and we need to set realistic goals for improvement. We need to use test scores in conjunction with a wide variety of other measures, and we need to balance the incentives to raise scores. We need to take steps to reduce inappropriate test prep.

We must *stop pretending that one test can do everything*. It's now common to claim that a test designed and used for accountability can also provide honest monitoring of progress and good diagnostic information for teachers. The fact that some are making this claim is hardly surprising; accountability testing has already swallowed a great deal of school time, and with our current incentives, few people want a second measure that might distract from the all-important goal of ratcheting up scores on the accountability test. However, it just isn't so, particularly given the pressures in our system to raise scores.

We need to use test scores in conjunction with a wide variety of other measures, and we need to balance the incentives to raise scores.

A corollary is that we need to *curtail sharply the use of the "interim" or "benchmark" assessments* that are widely used to predict how students will score at the end of the year. Many of these tests are just facsimiles of parts of the end-of-year summative test, designed to mirror not only the content of the summative test but also how that content is presented. Currently, students in many districts spend a huge amount of time over the course of the school year taking them. This is a waste of instructional time, and it is a recipe for score inflation. Obviously, tests used during the course of the year should reflect the same curriculum—the same domain—as the summative test, but they shouldn't be mirror images. They shouldn't be test prep.

Finally, a recommendation for a truly fundamental shift: we should consider turning the current approach on its head and *treating scores as the starting point rather than the end of evaluation*. I've stressed repeatedly that scores alone, whether high or low, aren't enough to tell us why students are performing as they do. Low scores, however, are an indication of likely problems. Rather than treating these low scores as sufficient to label a school a failure, we could use them to target other resources used for evaluation.

*For more on what international comparisons can tell us about American education, see "Puzzling Out PISA" in the Spring 2015 issue of *American Educator*, available at www.aft.org/ae/spring2015/schmidt_burroughs.

Provide Support to Teachers

Teachers can't do it all—especially teachers in many low-performing schools. This fact is widely accepted in principle, but it is often ignored in practice. We will need to take this far more seriously than we have if we are to achieve the large gains in student learning and, in particular, the big improvements in equity that reformers have promised us for years.

The supports we should provide are of three types. The first is *better initial training and ongoing support* for teachers already in the workplace. Many teachers simply don't have the skills needed to produce the improvements we want, particularly for disadvantaged children. There is nothing new about this recommendation. For decades, American experts in teacher training, such as Linda Darling-Hammond of Stanford University, have been pointing to the need for better training and internships.*

The second category is *in-school supports*: supplementary classes, longer school days, smaller classes, and the like. The third is *out-of-school supports*; one that has received a great deal of attention in recent years is high-quality preschool, which can improve the long-term prospects of disadvantaged kids.

Why are recommendations for more support controversial? One reason is money. It is vastly cheaper to buy a test, set arbitrary targets, and pretend that the problem is solved. A second is timing. It takes time for these supports to work. Test scores can be improved very rapidly—even in the space of only two or three years—if one turns a blind eye to fraudulent gains.

There is one additional, less obvious reason why the importance of support might be controversial: its implications for setting targets. Just as the improvements we can reasonably expect depend on the circumstances confronting any given school, they also depend on the amount of support we are willing to provide to the educators who work in it.

For example, consider two hypothetical elementary schools that are located in very poor neighborhoods and that largely serve highly disadvantaged students. Assume that the teachers in the two schools are comparable in quality. Students in the first school have access to high-quality preschools, health screening, and a school breakfast program. The second school has none of these. It would be unrealistic to expect students in schools like the second to match the performance of kids in schools like the first, and expecting similar performance would necessarily cause you to conclude—falsely—that teaching in the second school is of lower quality. Once again, this points to the importance of knowing about the context in which a school operates and to the need for professional judgment.

In this brief article, I could only describe some of the steps we should take to replace test-based accountability with something more effective. I couldn't describe in detail the failures of test-based accountability or the principles underlying the alternatives I recommend here. I discuss these in depth in *The Testing Charade*.

Implementing these recommendations will be a daunting task. To start, it will require a great deal more work than simply testing students. Even if well designed, a new system will also require patience; the obstacles to improvement are substantial, and nothing will produce gains as rapid as the bogus gains in scores we have become accustomed to with test-based accountability. And a better system is likely to be considerably more expensive—if one doesn't count the cost of the countless hours of potential instructional time we are now tossing away for test prep and excessive testing.

And we need to face up to two basic facts about interventions in complex systems such as education: most interventions, even very good ones, will have side effects we don't want, and none will work exactly as planned. The implications of this are clear. We need to monitor—routinely—the effects of any new interventions, and we need to be prepared to face the music and make midcourse corrections when warranted. We expect this in fields like medicine and auto safety, and we ought to demand it in education as well.

No matter how large, however, these difficulties don't provide an excuse to continue on the current path. The strategy of test-based accountability has failed, and tinkering around the edges won't change that. Everyone with a stake in our educational system—including parents, employers, educators, and most importantly students—deserves better. ☐

*For more on teacher supports, see "One Piece of the Whole" in the Spring 2014 issue of *American Educator*, available at www.aft.org/ae/spring2014/darling-hammond.

Endnotes

1. "California Lawmakers Target Linking of Student IDs to Test Scores," *On Politics in the Golden State* (blog), *Los Angeles Times*, June 18, 2012, http://latimesblogs.latimes.com/california-politics/2012/06/student-notebooks-standardized-testing.html.

2. Jessica Chasmar, "Common Core Testing Makes Children Vomit, Wet Their Pants: N.Y. Principals," *Washington Times*, November 25, 2013. See also Katrina vanden Heuvel, "Stakes on Standardized Testing Are Too High," *Washington Post*, April 30, 2013.

3. Stephen Klein et al., *What Do Test Scores in Texas Tell Us* (Santa Monica, CA: RAND Corporation, 2000); Diana Jean Schemo, "Questions on Data Cloud Luster of Houston Schools," *New York Times*, July 11, 2003; Diana Jean Schemo and Ford Fessenden, "Gains in Houston Schools: How Real Are They?," *New York Times*, December 3, 2003; Andrew Ho and Edward Haertel, *Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples* (Los Angeles: Center for the Study of Evaluation, 2006); and Robin Tepper Jacob, Susan Stone, and Melissa Roderick, *Ending Social Promotion: The Response of Teachers and Students* (Chicago: Consortium on Chicago School Research, 2004).

4. Everett F. Lindquist, "Preliminary Considerations in Objective Test Construction," in *Educational Measurement*, ed. Everett F. Lindquist (Washington, DC: American Council on Education, 1951), 119–184.

5. C. Kirabo Jackson, "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes," NBER Working Paper Series, no. 22226 (Cambridge, MA: National Bureau of Economic Research, 2016).