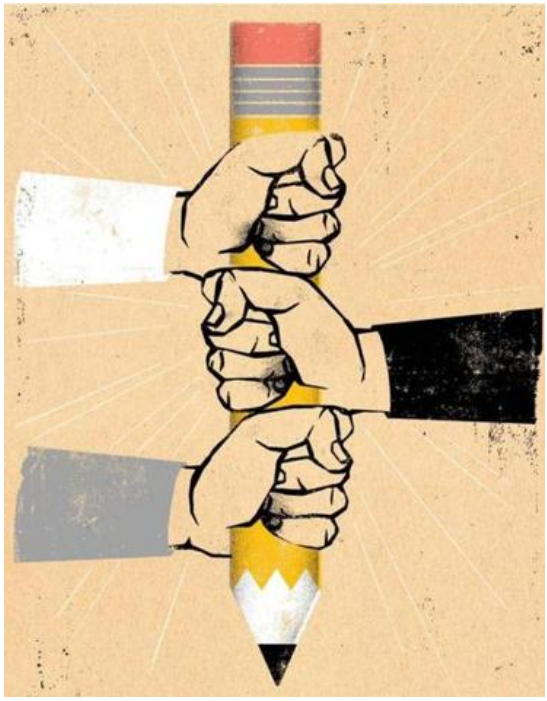
The Boston Globe

Ideas

What to test instead

A new wave of test designers believe they can measure creativity, problem solving, and collaboration – and that a smarter exam could change education.

By Leon Neyfakh | GLOBE STAFF SEPTEMBER 16, 2012



EDEL RODRIGUEZ FOR THE BOSTON GLOBE

When Harvard University announced last month that it was investigating 125 students for cheating on a take-home exam, most of the ensuing public fuss focused on the students: whether they were kids wrongfooted by the requirements of an unpredictable class, as they claimed, or sneaky overachievers driven to cut corners by some mix of ambition and laziness. But beyond the question of the moral fiber of Harvard students, there was another player in the drama: the test itself.

The final exam, according to the instructions, had been "completely open book, open note, open Internet, etc." The one thing it forbade the students to do was to work together—a requirement that some commentators, such as Slate columnist Farhad Manjoo, have argued is absurd. If the purpose of education is to help young people develop skills they'll need later in life, Manjoo wrote after the scandal broke, it makes no sense to arbitrarily prevent them from demonstrating precisely those skills when they're taking a test.

That point doesn't exonerate the accused: If there was a rule and they broke it, they

cheated. But it does raise a deeper question: Just what was the test trying to achieve? What exactly do we want our tests to be testing?

Being successful in today's world, as we all now recognize, requires more than an ability to think quickly and recall facts on command. And our education system has, however fitfully, moved to address those values. The problem is that our tests still lag behind. A final exam like the Harvard one, for example, attempts to test students' resourcefulness by allowing them to consult books and websites—but then, because professors still need to grade on individual performance, draws a hard line on working in groups, which in the real world is an important part of resourcefulness as well. That students were so tempted to use these illicit skills of collaboration points at the problem: There is a whole spectrum of crucial skills—creative thinking, problem-solving, communicating with others—that educators are still struggling to test for in a fair, objective, efficient way.

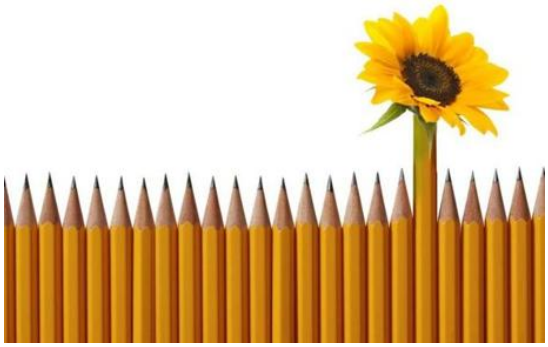
Reengineering tests has become a kind of calling for a group of educators and researchers around the country. With millions of dollars of funding from the federal government, the Bill and Melinda Gates Foundation, as well as from firms like Cisco Systems, Intel, and Microsoft, they have set about rethinking what a test can do, what it can look like, and what qualities it can assess. Using new testing ideas like computer simulations, games, and stealth monitoring, they are trying to take what they believe is a huge and necessary leap—changing the test as we know it from a fixed measurement of what a student can remember on a particular day, to something far more dynamic and informative. If the research pays off, college students of the future may find themselves taking tests that those of us who remember blue books and scantrons won't recognize as tests at all.

The researchers at the forefront of test design also have a bigger dream, rooted in the idea that tests aren't just a static part of education, but can actively shape what teachers teach and what students learn. If you can really build smarter, more sophisticated tests, they say, you can change education itself.

"[Tests are] the tail that wags the dog," said David Williamson Shaffer, a professor at the University of Wisconsin-Madison who studies education psychology. "And the problem is we've got the wrong tail on right now. We have a tail that was literally developed 100 years ago."

***

When most of us imagine taking a test, we picture a stressful, high-stakes trial in which a teacher or an employer puts us on the spot to find out how good we are and how much we know. Tests determine our grades, ranks, and qualifications; they are one of the main tools society uses to determine who goes where and who can do what.



GLOBE STAFF PHOTOILLUSTRATION

But those whose job it is to design tests understand them as something else: a guess about the future. When we test, we're really probing for certain qualities—the particular mix of knowledge and ability—that tell us a student is ready to move ahead, or an employee will be an asset to the firm.

Such predictions require a clear sense of the qualities a person needs in order to thrive. Over the past several decades, educators have changed their minds about what those qualities are, as skills like collaboration have become more important in the workforce. "The nature of what the world is calling for is changing," said Randy Bennett, a research scientist at the testing company ETS. Added his colleague, Robert Mislevy: "There are just a lot fewer jobs where you're not doing information-seeking, interpreting, problem-solving, and communication than in the past."

When it comes to measuring those things, it's obvious to researchers that traditional written tests—even ones that are more free-form and open-book—are not up to the task. "You think of true/false, multiple choice, maybe matching and maybe some fill in the blank and perhaps an essay question," said Valerie Shute, a professor of education psychology and learning systems at Florida State University. "Those kinds of things are extraordinarily narrow as far as the scope of what you can get."

To broaden the scope, researchers like Shute are trying to engineer tests—they prefer the term "assessments"—that require people to exercise a bundle of complex skills at the same time, not just regurgitating information but using it to solve problems. Kathleen Scalise, an associate professor at the University of Oregon who studies how computers can be used in learning, has mapped out a taxonomy of testing innovations that includes a range of nontraditional types of questions. One asks students to move a pair of street lights around so that a woman shown on the screen casts two shadows; another shows 15 bubbles containing words like "Congressmen," "President,"

“Supreme Court,” and “Justices,” and asks students to connect them to each other using arrows and arrange them in clusters. At the progressive New York school Quest to Learn, students are put into small groups and instructed to build Rube Goldberg machines to demonstrate an understanding of basic physics and an ability to work in teams.

But researchers in the field of test design are pouring the most energy into crafting computer programs that take advantage of so-called stealth assessment, a method of judging test-takers without telling them exactly what’s being judged. Jody Clarke-Midura and Chris Dede at Harvard, for instance, developed a 3D video game to test scientific skills: In one assessment, farmers in a village find a dead frog with six legs, and students must help them figure out the cause of the aberration by walking from farm to farm gathering data. To test students’ grasp of physics, meanwhile, Valerie Shute and her team designed a computer game called Newton’s Playground. The program gives students a problem and asks them to use their mouse to draw tools, such as pendulums, ramps, and levers, in order to solve it. Getting the answer right matters, but by watching how kids go about trying to solve the problems, the test can also evaluate whether they’ve learned important concepts like inertia and momentum, as well as their ability to be persistent and think creatively. The test, in other words, opens a window into how the student’s mind actually works.

The key conceptual advance embedded in stealth tests like Newton’s Playground is that students are no longer being evaluated merely on their final answers, but the process they go through to attack a problem. That’s also the principle that drives the work of David Shaffer, who has built a multiplayer game about designing a city that grades students on their collaboration skills by monitoring and recording their interactions. “It tracks everything they do,” Shaffer said. “Every e-mail they send, every time they communicate with someone through the chat system, every report that they write, every page that they look at....That’s an awful lot of information we can use to figure out how well they’re doing what they’re doing.”

But if Shaffer and other next-generation test designers share a dream of replacing pen-and-paper exams with process-oriented problem-solving exercises, they also share a thorny challenge: The skills they’re trying to measure are much harder to detect and quantify than, say, whether someone knows the quadratic formula. “It’s not just that [complex skills] are harder to isolate—it’s that they don’t exist in isolation,” said Shaffer.

Breaking down these multifaceted skills into testable qualities is difficult, and it's something educators have been trying and failing to do for more than half a century. The first president of ETS, which has long administered the SAT, set out in 1948 to develop a test that could evaluate a student's intellectual stamina, ability to get along with others, and so on—but the company eventually concluded it was too hard to measure in a reliable way. More recently, in the late 1980s and '90s, the Harvard developmental psychologist Howard Gardner participated in an effort to design new kinds of tests in the humanities that could be graded objectively. Ultimately, he found that the nuance required to measure softer skills collided with the demands of standardization. When a test needs to reliably compare students from across schools and districts, "there is pressure to simplify, have ironclad rubrics, essentially move toward multiple choice," Gardner wrote in an e-mail.

Technology has opened the door to loosening those demands, as computers are now able to track everything a person does while trying to solve a problem and translate it into numbers. But that still leaves open the trickier question of what mix of e-mails and keystrokes and clicks tells us that someone is, say, a good collaborator, and what constitutes a warning sign. Test designers have a few ways of trying to figure that out. One involves having groups of novices and experts complete a simulation, and building a grading system based on how their approaches differ; another involves matching performances on an experimental test with a long, time-intensive personal assessment. If that sounds difficult and expensive—well, there's a reason multiple choice has lasted so long.

"[Historically], the testing industry, because it was pragmatic, only tested what it was easy to test," said James Paul Gee, a professor at Arizona State University who designs tests that take the form of games. "But as a parent, I don't want you to just test what's easy to test, I want you to test what's important to test."

***

As we grow up, the tests we take shape our understanding of what matters in the world. You know checking your blind spot when switching lanes is important in part because if you don't do it, the DMV guy won't pass you on your driver's exam. Or consider the MCAS, which high school students in Massachusetts must pass to graduate: It covers only English, math, and science, reflecting—and ultimately enforcing—the back-to-basics belief that being proficient in those areas is the bedrock

of an education, while a grasp of history, civics, or arts is not.

In a very direct way, then, a test expresses a set of values—a vision of how we want people to be. “A test becomes a sign post,” said Randy Bennett. “It becomes an example of what to strive for.”

This is what motivates many test designers, whose ultimate goal is to expand what is taught in schools by expanding what can be tested. If we had better tests, they imagine—ones that could really assess abilities like creativity and collaboration—the door would suddenly open to a vision of education in which seemingly “soft” skills could be taught just as rigorously as what we consider the basics.

Things being as they are, we bristle at the idea of teachers feeling pressured to “teach to the test,” out of fear their students will otherwise score poorly on nationally mandated standardized tests. But what if teaching to the test didn’t have to mean mechanically training kids to crank out answers to invented questions? That’s the promise of a better test: By drawing a map that more accurately reflects our world, we may discover far more promising paths to get where we want to go.

Leon Neyfakh is the staff writer for Ideas. E-mail lneyfakh@globe.com.