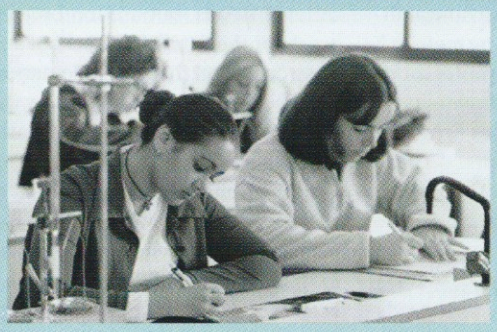
REINVENTING The BIG TEST

Problem: Old-school accountability tests are crude measurements of student learning.
Solution: Build a better test.

By Grace Rubenstein Photography by Gregory Cherin



past

Traditional standardized tests deaden teaching and inaccurately measure student learning.

future

Desiree Jerome demonstrates a chemiluminescent reaction as part of her portfolio to move up to the next grade at F.W. Parker Essential School, in Devens, Massachusetts.

When I was a younger education reporter in the old mill town of Lawrence, Massachusetts, the big day came when the state released scores on its school accountability tests. The Massachusetts Comprehensive Assessment System, better known and feared as the MCAS, fulfills the requirements of the federal No Child Left Behind Act through annual tests in English and math (and now additional subjects).

I scrutinized pages of numbers and wrote a story on the success and failure of nearby schools. My editors played it big on the front page because they knew parents would look anxiously at their school's results and homeowners would mentally adjust their property values based on the scores. I prodded principals and superintendents to explain their schools' leaps or stumbles.

And unwittingly, I played right into the dominant illusion that these bloodless test scores are the most definitive measure of a school's success—and that they measure what's most important.

Cold, hard numbers have a way of seeming authoritative, but accountability tests are not the infallible and insightful report cards we (and our state governments) imagine them to be. The educational assessment tests that states use today have two fundamental flaws: They encourage the sort of mind-numbing drill-and-kill teaching that educators (and students) despise and, just as importantly, they don't tell us much about the quality of student learning.

Evaluation Nation

Visit Edutopia.org *for many more feature articles, expert interviews, and video documentaries about assessment, including these:*

poll

TESTIFY ABOUT TESTS

Vote about which skill standardized tests should most emphasize at
www.edutopia.org/poll-test-skill

features

ACCURATE ASSESSMENT

Learn about a school where grades mean something at
www.edutopia.org/assessment-scoring-rubrics

F FOR ASSESSMENT

Read a discussion of the failure of standardized testing at
www.edutopia.org/assessment-flaws

HEALTHIER TESTING MADE EASY

Learn about authentic assessment at
www.edutopia.org/authentic-assessment-feedback

STUDIES IN SUCCESS

Check out a survey of research about assessment at
www.edutopia.org/assessment-research

interviews

EVALUATION EXPERTS SPEAK

Renowned educators discuss the ramifications of high-stakes testing at
www.edutopia.org/high-stakes-testing

HOWARD GARDNER

The multiple-intelligences pioneer sounds off on assessment at
www.edutopia.org/howard-gardner-interview

TESTING WITH TECH

Read about the role of technology in supporting and enhancing assessment at
www.edutopia.org/technology-assessment

videos

ASSESSMENT OVERVIEW

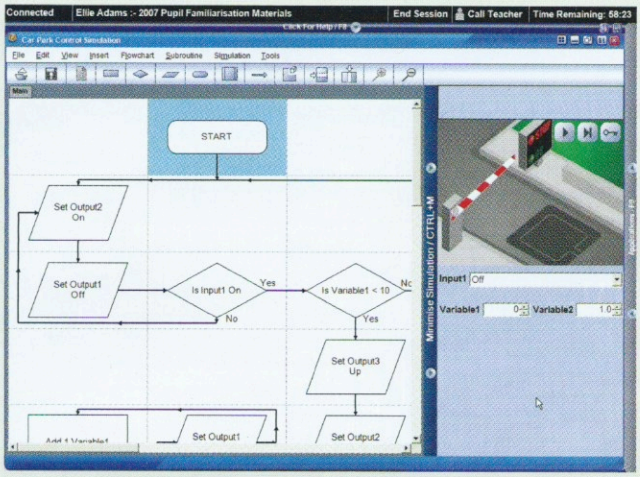
Note how performance assessments offer a richer, more holistic approach to evaluating what students know and can do at
www.edutopia.org/assessment-overview-video

URBAN ACADEMY

See how a small New York City high school employs project-oriented assessment at
www.edutopia.org/urban-academy-testing-video

STUDENT BUILDERS

Watch what happens when real architects judge a student project to design a school for 2050 at
www.edutopia.org/mountlake-terrace-video



This ICT Literacy Test item asks students to write a program controlling the operation of a parking-garage gate.

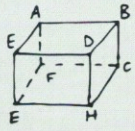
Tech Literacy, the British Way

The British government has tackled head-on the need to cultivate one essential twenty-first-century skill: computer literacy.

This year, U.K. schools began using the ICT Literacy Test for students ages 11–14 to gauge not only their mastery of technical skills but also their readiness to apply these skills effectively in everyday life and work.

Far beyond the simple keyboarding tests of old, this exam challenges students to create presentations with text and images, manipulate databases, and write simple computer programming, among other skills. Basic techniques such as saving information, using email, and doing simple searches are included, too. The test, taken entirely on a computer, embeds these assignments in practical tasks, all done in the virtual town of Pepford.

Sue Walton, project director at the United Kingdom's National Assessment Agency, an arm of the Qualifications and Curriculum Authority (QCA), explains that the emphasis is on students "actually being asked to do things."



To begin, a student might receive an email from the director of the local visitor center assigning him to design a tourist brochure, or from the Pepford High School principal requesting that she assess the effectiveness of a recent campaign to promote eating fruit instead of candy. Then the student would use information, charts, photographs, and other resources available within the virtual Pepford world to solve the problem. Test makers designed a full set of generic software—an email program, a Web browser, a database manager, and more—to avoid endorsing any one commercial brand or favoring students who are already familiar with certain programs.

Students' responses are scored dynamically, meaning that the computer captures the process they use to answer a question. For instance, if the test asks pupils to use a database to figure how many musicians play rock music, they could do this simply by counting or by using the filter, sort, or query tools. The computer gives students credit for a right answer while also evaluating their process and producing an instant report on how basic or advanced their skills are.

By the end of 2008, a battery of fifteen- to thirty-minute tasks will be available to teachers on demand, anytime. The test is not mandatory, but it's free, and Walton expects most schools to use it to help tailor instruction.

Creating a test like this demands investment of time and money: All told, the QCA put about $46 million into this six-year project.
—GR

"We are totally for accountability, but we've got the wrong metrics," says John Bransford, a professor of education at Seattle's University of Washington who studies learning and designs assessments. "These tests are the biggest bottleneck to education reform."

Hobbled by History

Jennifer Simone, a fifth-grade teacher at Deerfield Elementary School, in Edgewood, Maryland, is acutely aware of the limitations of standardized tests. Her curriculum must emphasize subjects for which the state accountability test measures proficiency—math, reading, and science. Social studies? Though the subject is on her master schedule, if there is a shortened school day, it gets dropped.

Moreover, Simone says, the test scores don't truly reflect her students' abilities and are too vague to help her pinpoint individual needs. She longs for an assessment that relies on more than just written problems, that could capture the more diverse skills visible in her classroom and valued in the workplace, such as artistic talent, computer savvy, and the know-how to diagnose and fix problems with mechanical devices. Simone asks, "If we differentiate our instruction to meet the needs of all the learners, why aren't we differentiating the test?"

The simple, but unsatisfying, answer is history and efficiency. The tests that states use to satisfy NCLB descended from a model created in the 1920s designed to divide students into ability groups for more efficient tracking. Eighty years, two world wars, and a technological revolution (or two) later, the tests remain structurally the same.

Policy makers revere the seeming objectivity of these tests, but the truth is the exams are not adept at determining either how well teachers have taught or students have learned—and test makers themselves will tell you so. Stephen Dunbar, an author of the influential Iowa Test of Basic Skills, explains that these tests can help illuminate statewide educational trends, but are too broad a brush for the detail at the school and classroom level that NCLB demands.

Assessment tests might show the overall effectiveness of the ninth-grade curriculum, for instance, or indicate trends within large demographic groups in that grade. But Dunbar says that when you get down to measuring the ability of students at Dallas's Woodrow Wilson High School, for example, where you're comparing this year's ninth graders to last year's, accountability test scores are not very useful. "They might tell you more about idiosyncrasies in that combination of kids than the level of achievement or the quality of teaching and learning that's going on," Dunbar explains.

In other words, state governments, at the behest of the feds, are using tests to measure something they actually don't measure very well, and then penalizing schools—and in some cases, denying students diplomas—based on the results.

"Most of these policy makers are dirt ignorant regarding what these tests should and should not be used for," W. James Popham, professor emeritus at the University of California at Los Angeles and former president of the American Educational Research Association, told PBS's *Frontline* in 2001. "And the tragedy is that they set up a system in which the primary indicator of educational quality is simply wrong."

There are several reasons the tests are imprecise (see "Where Standardized Tests Fail," page 37). Some are technical: an ambiguous question, a misjudgment in setting the difficulty level, a scoring error. The National Board on Educational Testing and Public Policy, at Boston College, has documented cases when scoring errors sentenced children to summer school or caused them to miss graduation before the mistakes were discovered. Some reasons are personal: Simone, whose school narrowly dodged state intervention last year, has seen fifth graders arrive on testing day angry about personal matters; others struggled to sit still during the test or

broke down in tears under the pressure.

The tests' fallibility has most to do with the very idea of measuring a year's worth of learning in a single exam. Inevitably, cramming that much coverage into a short test leads states to rely mostly on multiple-choice questions—the fastest and cheapest means of large-scale assessment. Such brief yet weighty exams limit the ways students can show their skills, and because it's impossible to test hundreds of state standards in a few hours, they leave teachers guessing on which to emphasize. Randy Bennett, who holds the title of distinguished scientist at ETS, writes that this rigid idea of assessment yields a "narrow view of proficiency" defined by "skills needed to succeed on relatively short, and quite artificial, items."

Even when states do pony up to use open-ended essay questions and pay human scorers, these questions can encourage formulaic answers. Last school year, I watched the principal of a (high-scoring) Boston high school interrupt a test-prep session to warn students not to stray from the essay-writing formula—main idea, evidence, analysis, linking—lest they lose points. "Don't be creative," she said fiercely. "You've heard me rail against standardized tests, and this is why. There's one way to do this, and it's the way the assessment coordinator told you."

Equally worrisome is that today's assessments emphasize narrow skill sets such as geometry and grammar, and omit huge chunks of what educators and business leaders say is essential for modern students to learn: creative thinking, problem solving, cooperative teamwork, technological literacy, and self-direction. Yet because NCLB has made accountability tests the tail that wags the dog of the whole education system—threatening remediation and state takeover for schools that fall short—what's not tested often isn't taught.

In short, the American accountability system is a bastion of the past that's stifling our ability to tackle the future.

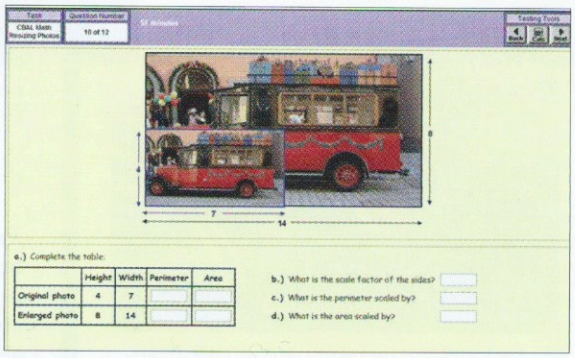
High Stakes

The good news is there's work afoot to create better tests that will challenge students to demonstrate more creative, adaptable skills—and, in turn, encourage teachers to teach them. Some model assessments already exist; for instance, many experts tout the Programme for International Student Assessment (PISA) exam for its challenging, open-ended questions on practical topics, such as climate change or the pros and cons of graffiti. Even more advanced models, some using computer simulations, will become available in a few years—and none too soon.

Business leaders have issued dire warnings about how hard the U.S. economy will tank if our education system doesn't get itself out of the nineteenth century, and fast. They're clamoring for creative, productive, affable employees—not just dutiful test takers—and they point to assessment as a crucial tool for turning the tide. Microsoft founder Bill Gates, addressing state governors, CEOs, and educators at the National Education Summit on High Schools in 2005, said, "America's high schools are obsolete. Even when they're working exactly as designed, they cannot teach our kids what they need to know today. In the international competition to have the biggest and best supply of knowledge workers, America is falling behind."

The New Commission on the Skills of the American Workforce, convened by the nonprofit National Center on Education and the Economy, issued a stark report in December 2006 predicting that our standard of living "will steadily fall" compared to other nations unless we change course. The globalized economy has created, the commission wrote, "a world in which comfort with ideas and abstractions is the passport to a good job"; what's essential, it added, is "a deep vein of creativity that is constantly renewing itself." According to the report, whatever efforts we make to modernize education, without a complete overhaul of the testing system, "nothing else will matter."

Congressman George Miller, chairman of the House Education and Labor Committee and chief House wrangler of NCLB (and a member of The George Lucas Educational Foundation's Advisory Board), understands the problem. The original law left it up to states to choose their own tests, but now he believes most states picked tests more for



This ETS test item taps math skills by asking students to properly resize digital photos.

Ending Hit-and-Run Testing

Strange as it might sound, a big push to reinvent standardized tests is coming from a major standardized-testing company, ETS.

The Princeton, New Jersey, nonprofit organization, which produces the SAT and Advanced Placement exams, among others, is two years into a five- to ten-year project to create an accountability test that—unlike the tests states use today—measures complex, real-world skills and helps teachers improve instruction.

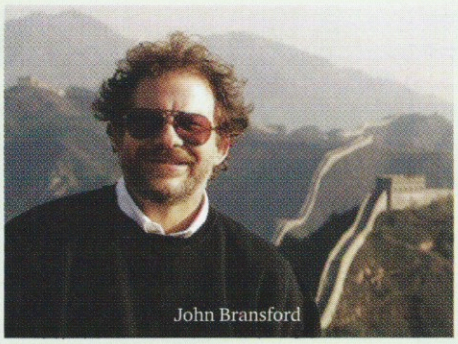
"What we are trying to do is come up with tests that not only measure discrete skills but also measure their integration," says ETS distinguished scientist Randy Bennett, "tests that exemplify not only the kinds of things that students must know and be able to do to succeed in the twenty-first-century world but also the kinds of things that teachers want to teach."

The ETS vision is to create a far longer assessment than today's quick-hit exams, then break that assessment up into many parts that could be done in short sessions over the course of a whole school year. The added time would allow test makers to use open-ended tasks that call on multiple skills, and place the tasks in meaningful contexts. For example, one task being developed calls on students to show their knowledge of mathematical proportion by resizing digital photos and explaining why certain sizes will or won't work. If test makers get it right, Bennett says, the exam should be a learning experience in and of itself, not an endgame.

Amassing test results over time, rather than at a single sitting, would prevent fleeting disturbances such as an argument or a poorly air-conditioned room from skewing kids' final scores. It also would provide teachers with feedback on student progress throughout the year—when they can actually use it—giving a richer, more reliable picture of student skills.

The hope is that what ETS calls the Cognitively-Based Assessment Of, For, and As Learning, or something like it, could ultimately replace the worn-out exams now being used to satisfy the No Child Left Behind Act. It's an uphill battle; it will take time, money, political will, and probably more advanced artificial intelligence systems to score written answers without breaking the bank.

But, as Bennett says, "I haven't heard anybody say, 'Don't try." *—GR*


John Bransford

Forget the facts. Can you learn?

Researchers at Seattle's University of Washington are creating a new kind of assessment that would turn our age-old ideas about learning on their head.

Contrary to popular belief, says project leader John Bransford, learning basic facts is not a prerequisite for creative thinking and problem solving—it's the other way around. Once you grasp the big concepts around a subject, good thinking will lead you to the important facts. So, along with colleagues at the university's Learning in Informal and Formal Environments (LIFE) Center, a laboratory sponsored by the National Science Foundation, Bransford is building assessments of what he calls "preparation for future learning."

"What we want to assess is how well prepared people are to learn new things in a nonsequestered environment where they have access to technology tools and social networks," says Bransford. Compared to typical standardized tests, for which seeking new information would be considered cheating, he says this model is "way more motivating, much more interesting for students, and much more valid in terms of what people really need to do when they get out of school."

In the computer-based tests, students are presented with complex problems that might have more than one good solution. One test challenges students to assume the role of animal-endangerment expert, fielding questions from fictitious clients around the world about how to protect local species. Another makes them virtual genetic counselors, dispensing advice to couples about potential risks to their children. They need enough conceptual knowledge to decide what kinds of questions to ask; then they search the Web for information and create whatever charts or diagrams will help them meet the challenge. Scoring is done with rubrics.

Fully realized, this kind of assessment would be linked with curriculum. Rather than moving along a metaphorical conveyor belt from one lesson to the next, Bransford says, students would spend time developing expertise in a subject. Through repeated challenges, they'd build up strategies and resources over time, just as a worker would on the job.

The researchers are trying out the test now with students in North Carolina and Washington State; they aim to have a prototype science test ready by the end of this school year. *—GR*

cost and efficiency than for educational value. "They don't truly measure what a student knows or doesn't know," he says, "or whether students have a depth of understanding so that they can apply their knowledge."

Real Solutions to Real Problems

In the past, states haven't had much choice in the kinds of large-scale assessments available, nor have they asked for much. That's about to change.

Test makers in multiple corners are creating more complex assessments, ones that, if tied more closely to curriculum and instruction, could paint a clearer picture of student learning. They're building these assessments to measure the twenty-first-century skills we so urgently need, aiming to gauge a child's readiness for the real challenges that await. If tests like these succeed, they could not only provide better information about children's readiness for real life but also give educators incentive to do what they want to do anyway: teach kids in engaging ways to be well-rounded people and lifelong learners, not drill the life out of school with dry test preparation.

A number of researchers are building tests that could be models—or at least one piece of a larger model. John Bransford and Andreas Schleicher, head of the Indicators and Analysis Division at the Organisation for Economic Co-operation and Development (OECD), maker of the PISA exam, believe students need dynamic problems to solve, ones that require real-world research and allow them to learn on the spot, not just apply prior knowledge.

A static problem, for instance, would ask test takers to say from memory how to save a certain endangered bird species. A dynamic assessment (in a real example from Bransford's lab) asks students to use available resources to learn what it would take to prevent the white-eyed vireo from becoming endangered. This is a novel question that demands students independently dig for information and know enough to ask the right questions to reach a solution.

Bransford says he doesn't believe the old trope that students must master a battery of content-specific facts before they can have a prayer of learning higher-order skills. "Just the opposite," he says: Students need to understand big concepts in each discipline, such as the relationship between a species' life cycle and its risk of extinction, but from there it's the higher-order skills that lead them to the pertinent facts.

At ETS—which writes the SAT and Advanced Placement exams, among others, and administers fifty million tests a year—Randy Bennett is field-testing assessments that make use of about thirty years of psychology research on how children learn. It's research that he says has been largely left out of test design. The key strategies he has found include asking students to integrate multiple skills (such as reading and making comparisons) at once, presenting questions in meaningful contexts, and using a variety of information forms, such as text, diagrams, and symbols. Eva Baker, codirector of UCLA's National Center for Research on Evaluation, Standards, and Student Testing, proposes one more: Never have someone present a solution without explaining why they chose it.

It's not so different from the kind of assessment Jennifer Simone would like for her students. She'd like the exam to use more formats than just writing, including visual or spoken components. "You would have to take the time to have a student interview, allow students to have an oral response," she says. "That's how we teach them reading."

Technology is what will make this revolution possible. Already, computers have enabled Bransford, Baker, and others to create interactive questions, search environments where students can find new information, and simulations to make problems more engaging and real. These tools can record students' answers as well as their thought process: what kind of information they sought, how long they spent on each Web page, and where they might have gone off track.

The British government has created a computer-literacy test that challenges teens to solve realistic problems (how to control crowds at a soccer match, for instance) using online resources. The more sophisticated these tools become, and the more adeptly test makers use them, the better assessment will be.

So, progress is coming—in some cases, has arrived—but as the OECD's Andreas Schleicher says, "It's a long road, and we're at the beginning." The biggest hurdles are time and money (richer tests require more of both to design and administer), and that rarely tamable beast, politics. The next version of NCLB, due later this year, could pump federal money into pilot projects to help states create richer assessments, paired with richer curriculum—but only if that clause survives the political battle to come.

Stephen Dunbar, the Iowa test author, has doubts that more complex tests can be done on a large scale. Though the effort is worthy, he says, the cost and time to create and score open-ended questions, and make

them comparable from year to year, could make it too impractical. Scary as it might sound, artificial intelligence is likely to play a big role in the scoring of such exams. If the technology becomes sophisticated enough to handle answers to trickier problems, it could make better assessment more affordable.

The ETS's Randy Bennett, on the other hand, believes the prospects of building an assessment system to match the demands of the twenty-first century are "pretty good." The key is to convince states that it's practical, affordable, and clearly better than today's exams at providing meaningful information. At least one state, West Virginia, has begun asking the test makers it contracts to emphasize more modern problems and skills. Another hurdle will be for politicians to temper their devotion to multiple-choice questions and get comfortable with a little subjectivity. "For any assessment," Schleicher says, "you have to make a trade-off between objectivity and relevance."

Jennifer Simone, for one, is depending on forward-thinking test makers and policy makers to succeed—for the sake of her students, most of all. "That we are held accountable is a good thing. That we are doing something to measure the progress of our students is a good thing," she says. "I just disagree with the way it's being done." e

command performance

While schools wait for innovation in accountability testing, some are taking matters into their own hands, creating performance assessments that guide and strengthen teaching and learning. Typically, these assessments come in the form of portfolios and presentations—tasks that bear something in common with the kind of work students may ultimately do in college or in a job.

At Anzar High School, in San Juan Bautista, California, students must complete a series of exhibitions to graduate, each one including a research-based written piece and an oral presentation. The topics are of the students' own choosing, fashioned (with guidance from a teacher-adviser) to cover language arts, science, history, math, and service learning and postgraduate plans—areas typically combined into three cross-disciplinary exhibitions. Students work for a semester or more on each project, and a panel of jurists, including teachers, alumni, and community members, evaluates their performance.

"If things are going as intended, students are really passionate about their issue, which means they're getting to devote a whole class period to working on something they adore," says Principal Charlene McKowen, whose school serves 420 students from San Juan Bautista and Aromas, rural communities south of San Jose. "It's almost eerie once they get going. You just hear 'Click-click-click,' and it's pretty quiet."

On exhibition day last spring, presentations covered such diverse topics as "Is Prison an Effective Rehabilitation for Latino Males?" "How Do Pets Affect Health and Education?" and "What Materials Will Be Used in the Future of Surfboard Manufacture?" Marisol Garcia, a junior who had researched the merits and failings of prisons, faced three panelists: a teacher, an alumnus now working at an international staffing company, and a San Jose State University professor. She told them about her interview with a prison guard, and drew connections between the data she'd found and a memoir she'd read, *Always Running: La Vida Loca: Gang Days in L.A.*, by Luis J. Rodriguez.

The verdict: Garcia excelled in analyzing the book but needed more substance in her factual presentation, the jurists said. They gave her a 2 ("minimal pass") for the history component and a 3 ("outstanding effort all around") for language arts. Said Garcia, "You have to actually know what you're talking about. It takes a lot of time and effort."

These assessments don't take away the pain of state accountability tests, but they do steer instruction toward critical thinking and endow students with confidence and useful skills.

In other schools, says McKowen, "I would just see over and over again that students would go off to college and be afraid or feel like a fraud, because they'd learned how to play the game. We wanted to be sure that any student who graduated from here would know what they were capable of doing." —*GR*

where standardized tests fail

Today's standardized assessments can be useful for spotting big trends or gauging the effectiveness of state programs overall. However, when used in high-stakes accountability, as the sole indicator of an individual student's achievement or the quality of a single school or school district, these tests can be imprecise. Creating and scoring such tests is complex. Here are some of the steps in the testing process where subjectivity prevails and inaccuracies arise:

TEST MAKING

- **Content selection:** If the state sets too many standards, teachers won't be able to cover them all and will have to guess which are on the test. If test makers include too few questions on any given skill, the results may not truly show how well a student can perform it.
- **Ambiguous questions:** Particularly for multiple-choice questions, a child may be able to make a plausible, even creative, argument for choosing one of the "incorrect" answers, but the format doesn't allow the child to explain.
- **Setting the difficulty level:** This determination, typically based on educators' and officials' opinions, is naturally subjective. To select final questions, test makers often try them out on students, which works only insofar as the trial-run group accurately represents the students who will ultimately take the test.
- **Year-to-year comparison:** To prevent cheating, states typically ask test makers to create new questions every year. Test makers must then perform the tricky business of trying to ensure that the exams are equally difficult so that scores can be compared like apples to apples.

TEST TAKING

- **Test preparation:** The teaching of test-taking strategies may favor some students and keep their scores from reflecting what they actually know.
- **Distractions:** Whether internal or external, distractions such as test anxiety, personal problems, lack of sleep, a sick classmate, or a broken air conditioner can distort students' scores.

TEST SCORING

- **Mechanical or human error:** Mistakes may occur in setting the answer key, feeding answer sheets into scoring machines, marking answers right or wrong, or other steps in the process.
- **Cut scores:** These cutoff points for passing and advanced scores are based partly on educators' and officials' judgment, so they're subjective. Also, given the natural imprecision of scores explained in this chart, a student's score may fall below the cutoff point for failing even if she is knowledgeable enough to pass—and vice versa. —*GR*