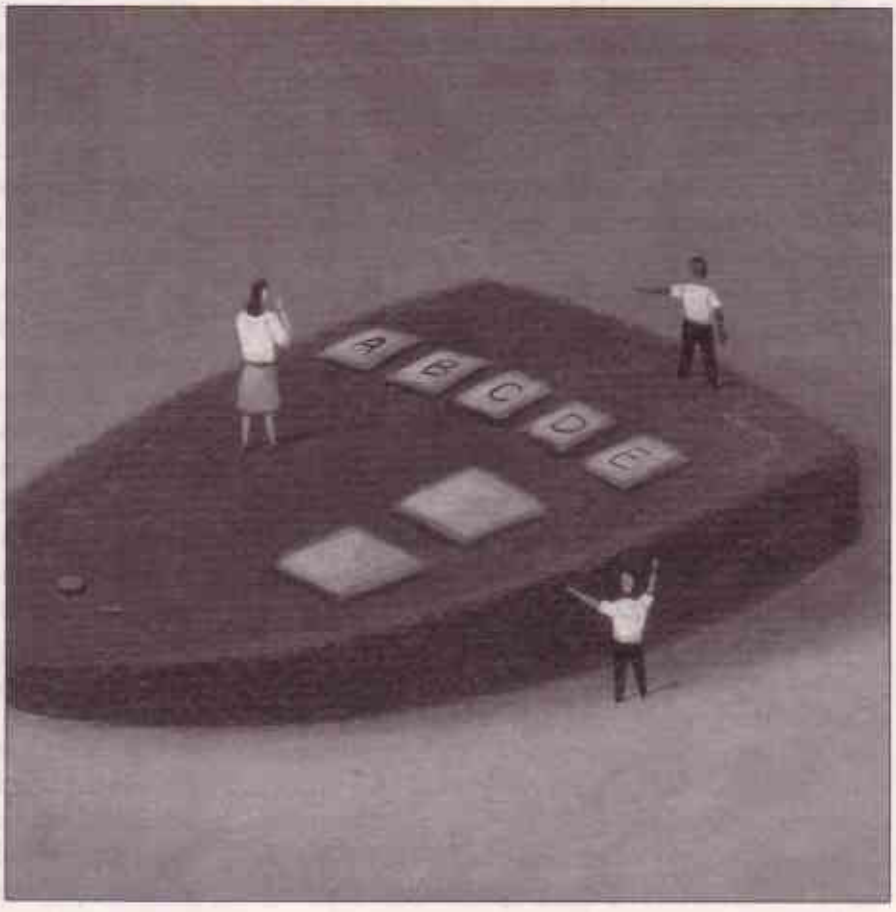
Alternative Approaches to High-Stakes Testing

Mr. Lederman and Mr. Burnstein propose a novel way to increase student engagement and counter the pressures of high-stakes testing.

BY LEON M. LEDERMAN AND RAY A. BURNSTEIN

UNDER No Child Left Behind (NCLB), the federal government requires state governors, superintendents, and school principals, through a regime of annual testing, to demonstrate to the taxpaying public that education dollars are being used effectively to improve student achievement. Developing, administering, and scoring the required assessments call for highly specialized skills and experience that states often lack, leading them to hire outside testing companies and consultants.

Student performance on these tests — given at the end of the year to all students in grades 3 through 8 — determines rewards and sanctions for schools, teachers, and students. But these standardized tests were not designed for accountability purposes, and experience in isolating and measuring the effects that schools have on student learning is rare. This test-based accountability system is being controlled by people who may know how to develop standardized achievement tests but know very little about the institutional realities of accountability — and even less



LEON M. LEDERMAN, Nobel Laureate, is resident scholar at the Illinois Mathematics and Science Academy, Aurora, and Pritzker Professor of Science at the Illinois Institute of Technology, Chicago. RAY A. BURNSTEIN is research and emeritus professor of physics at the Illinois Institute of Technology.

Illustration by John Berry

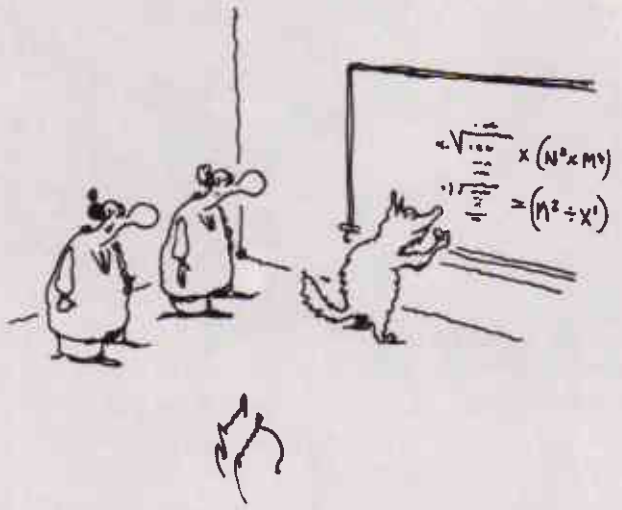
about how to improve instruction in schools. For the most part, the assessments currently in use are not capable of accounting for the many nonschool factors that influence test scores, including student background, home environment, poverty level, and English-language proficiency. Such standardized tests are not always reliable measures of what is learned in the classroom, according to assessment expert James Popham.[1] And now NCLB has established these single, end-of-year tests as the dominant measures of school success or failure.

HIGH-STAKES TESTING

Tests used to make high-stakes decisions, especially in light of the fact that they are given only once or twice a year, necessarily have to meet a set of strict requirements. High-stakes tests must be "instructionally sensitive" — that is, they must be capable of determining changes in achievement related to instructional improvements. The validity of an accountability system depends on designing the right tests. These tests should meet the following criteria:[2]

1. Assessments used to measure student mastery of specific content must include clear descriptions — brief, jargon-free, and teacher-friendly — of what is going to be assessed. Classroom teachers need these descriptions to understand in detail what is expected of their students.

2. Effective assessments focus on a modest number of significant curricular aims, drawn from content standards. The selected content standards clearly must be of major importance.



"Wait, I think Lassie may be trying to tell us something."

3. An instructionally sensitive test to be used for accountability purposes must report student performance in a way that enables teachers to know what aspects of their instruction need to be improved and what aspects are working well.

The tests currently being used to satisfy NCLB have been judged by the majority of teachers to be instructionally insensitive — incapable of measuring the effects of instruction on student performance. This can lead to tragic results because high-stakes tests can distort instruction and may encourage teachers to "teach to the test." Teaching to a bad test and spending months on drill and skill may boost scores but surely ends up turning off students. The NCLB mandates for AYP (adequate yearly progress) and public reporting of results put enormous pressure on students, teachers, principals, and superintendents to raise test scores. This pressure can lead, in extreme cases, to cheating. The more general response is for teachers to practice on past (and even future) versions of the tests and to restrict instruction to just those subjects that will be tested; this is known as item teaching.

These outcomes of high-stakes testing distort the traditional ideal of the teacher as one who makes every effort to achieve the goals of the curricula without regard to any particular test. The real blame for inappropriate forms of teaching in response to testing lies not with teachers but with state and national policy makers who create accountability systems centered on ever-higher test scores (AYP) with little regard for how these scores relate to better learning.

Accountability based on high-stakes standardized testing ignores the vast differences students bring into the schools. As teachers, we know how genius and creativity in students may be hidden from us by so many factors, including language, motivation, boredom, gender, and, perhaps most crucially, thinking that diverges from the textbook and from typical classroom instruction. High-stakes testing may not only turn off students but may also totally disconnect them from the learning process.

CLASSROOM ASSESSMENT AND NEW EDUCATIONAL TECHNOLOGY

The American Psychological Association's guidelines for test use specifically prohibit basing any judgment on a single test score.[3] This position recognizes margins of error and the need for multiple measures of a student's performance before making critical decisions.

It seems that a much more productive approach than NCLB's annual testing would be to integrate instruction and assessment. This is far from a new idea. When testing is an integral part of pedagogy, one is actually teaching to the test — no, rather, teaching *with* the test. As teachers, we are aware that testing, in the sense of raising questions to get students thinking, is an essential component of pedagogy. By embedding testing into the teaching process, we can try to ensure that students are thinking about the subject. An optimum marriage of questioning and explaining can enhance the learning process. This approach has a familiar name: classroom assessment.[4]

We believe that keypads combined with Internet technology can be used to achieve embedded assessment, day by day, even hour by hour, without imposing the deadly burden of high-stakes tests.

Educational research has defined two distinct types of assessment: summative and formative.[5] Formative assessment enhances instruction by deftly using questioning and quizzing to establish a feedback loop between students and teacher. An example of formative assessment, used in both lecture and laboratory, is "interactive engagement," in which a teacher leads students in activities that in some way yield timely feedback.[6] In contrast with formative assessment, summative assessment is similar in form and use — e.g., a final examination administered at the end of the semester or school year — to the high-stakes tests we have been criticizing.

Teachers can apply modest forms of technology to improve the use of formative assessment. In 1993, we at the Illinois Institute of Technology initiated the use of "keypads" during classroom lectures. Keypads are wireless electronic devices that enable students to respond immediately to multiple-choice questions that are projected onto a large screen throughout the course of a lecture. After about 30 seconds or at the instructor's discretion, the responses of all the students are compiled by a computer and presented in a histogram. Each individual student's response has also been recorded in the computer. We accidentally stumbled on this wireless electronic system — originally designed for interactive sales pitches — and modified its use for high school and college instruction based on actual classroom experience. We began making presentations on keypad-based instruction at meetings of the American Association of Physics Teachers in 1995 and have continued ever since.[7]

A decade ago, wireless keypad systems were available from only one source and cost $300 apiece. Now there are about five suppliers, selling hundreds of thousands of keypads per year (both radio frequency and infrared) to schools and universities at a fraction of the earlier price. The systems are often called electronic student response systems (ESRS), and their use represents an enhanced type of formative assessment.[8]

We propose using this technology to satisfy, in part, the new accountability requirements that have been imposed on schools in an attempt to address district, state, and federal concerns about the quality of education. We believe that keypads combined with Internet technology can be used to achieve embedded assessment, day by day, even hour by hour, without imposing the deadly burden of high-stakes tests. If this interactive student response system does, say, 70% of the job of assessing students' progress in grasping concepts and reaching understanding, a summative test could then be added in order to satisfy the accountability authorities with about one-third of the trauma we see in our current system.

We have, of course, nurtured and studied this technique in physics instruction only, but others have used wireless keypads for classes in English literature, biology, engineering, etc. The wide utility of ESRS should not be surprising since the technique is applicable to all subjects that can be assessed in part by multiple-choice questions.

HOW DOES KEYPAD-BASED INSTRUCTION WORK?

During a typical 40- to 75-minute high school or first-year college class, a teacher can interrupt six to 12 times to ask questions that are designed primarily to test students' grasp of the subject matter but also to generate discussion among students. Each question offers a choice of three to 10 possible answers. After a minute or so, the class results are presented as a histogram. If the histogram shows that most of the class missed the concept, the teacher has instant feedback and can take immediate steps to address the students' lack of understanding. One possible response to such feedback would be to encourage peer instruction, in which students discuss the question with their neigh-

bors for several minutes.[9] Bedlam! Perhaps, but remember, the students are arguing over subject matter. The teacher then asks the same keypad-quiz question again to check whether the concept has been clarified. Moreover, this technique creates the possibility that discussion and argumentation among students will become a habit — one that is practiced outside the classroom as well.

The tabulation of one semester's keypad-quiz grades may result in as many as 300 to 600 scores for each student. This is enough to give the teacher a very good evaluation of each student's status and progress. The keypad quizzes may, of course, be supplemented by one or two full-period tests.

Keypad-based questions currently follow the traditional multiple-choice format. Multiple-choice questions are not essays, but they can be given essay-type features. Since the keypad quizzes are computer graded, an item can have more than one correct answer or offer the option of correct and "almost correct" answers. For example, a teacher can ask students to select from several sentences the one that offers the best explanation of a concept. When a student chooses a next-best sentence along with the best, that second choice can be added to the student's grade for the question. This scheme is far from being perfected, but, as we gain more experience with the technology, we can develop more incisive questions that will both test and sharpen student understanding. In this way, we can hope to extend the range of questions to higher-level cognitive domains[10] than would otherwise be available with the multiple-choice format, thereby making keypad-based formative assessment an even more effective testing procedure.

KEYPAD-BASED ASSESSMENT AND ACCOUNTABILITY

High-stakes testing, even at its best, puts a strain on good pedagogy, places a huge burden on students and teachers, and creates winners and losers in an education system that needs to have all winners. Our federal and state education policy makers have not inspired confidence that their procedures can fix the current system. And we have not even mentioned the lack of sufficient funding for so massive a federal intervention as NCLB. As an alternative, classroom-embedded assessment can provide continuous, detailed information on the progress of students, and keypads and Internet technology can allow state and federal officials to augment their one-test approach.

With reasonable coordination between teachers, schools, and, say, accountability headquarters at the state and federal levels, an accountability system that combined keypad-based formative assessment with summative assessment could be created. State or federal education experts could develop standardized multiple-choice accountability tests and require students to take them on a semester or annual basis. Teachers could download the tests via the Internet and administer them at the correct phase of the class. Such a system would no longer rely on single high-stakes tests since the keypad data collected by the teacher would contribute significantly to the overall assessment of student achievement and in a different way from the summative test.

As educational technology becomes an increasingly dominant factor in pre-K–12 education, we need to be creative in looking at how it can modify curricula, assist the student, ease the administrative burden, and support the teacher in providing exemplary and joyful instruction.

1. W. James Popham, "Standardized Achievement Tests: Misnamed and Misleading," *Education Week*, 19 September 2001, p. 46.

2. See, for example, W. James Popham, "The Trouble with Testing: Why Standards Based Assessment Doesn't Measure Up," *American School Board Journal*, February 2003, pp. 14-17.

3. The American Psychological Association's testing guidelines are available at www.apa.org/pubinfo/testing.html.

4. Thomas A. Angelo and K. Patricia Cross, *Classroom Assessment Techniques: A Handbook for College Teachers* (San Francisco: Jossey-Bass, 1993).

5. Paul Black and Dylan Wiliam, "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan*, October 1998, pp. 139-49.

6. Richard R. Hake, "Interactive Engagement vs. Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, January 1998, pp. 64-74.

7. Ray A. Burnstein and Leon M. Lederman, "Interactive Lectures: Keeping Students Involved in a Lecture Course," *AAPT Announcer*, July 1995, p. 80; and idem, "Report on Progress in Using a Wireless Keypad Response System," in Edward F. Redish and John S. Rigden, eds., *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education* (College Park, Md.: American Institute of Physics, Conference Proceedings No. 399, October 1997), pp. 531-37.

8. For more information on the pedagogical value of ESRS, see Ray A. Burnstein and Leon M. Lederman, "Using Wireless Keypads in Lecture Classes," *Physics Teacher*, January 2001, pp. 8-11; and H. Arthur Woods and Charles Chiu, "Wireless Response Technology in College Classrooms," *The Technology Source*, September/October 2003, available at http://ts.mivu.org/default.asp?show=article&id=1034.

9. Eric Mazur, *Peer Instruction: A User's Manual* (Upper Saddle River, N.J.: Prentice-Hall, 1996).

10. Benjamin S. Bloom, *Taxonomy of Educational Objectives* (New York: Longmans, Green, 1956).