

## EDUCATION WEEK

Published Online: June 7, 2011

Published in Print: June 8, 2011, as **Panel Finds Few Learning Benefits in High-Stakes Exams**

# Panel Finds Few Learning Benefits in High-Stakes Exams

**Report by blue-ribbon commission likely to feed debate on the wisdom of tying consequences to students' test scores**

By **Sarah D. Sparks**

As Congress debates how to structure the next iteration of federal school accountability, a new national **study** has raised serious concerns about the effectiveness of test-based incentives to improve education.

[← Back to Story](#)

A blue-ribbon committee of the National Academies' National Research Council undertook a nearly decade-long study of test-based incentive systems, including the "adequate yearly progress" measures under the No Child Left Behind Act, high school exit exams, teacher merit-pay programs, and other testing-and-accountability initiatives. While the panel says it supports evaluating education systems and holding them accountable, on the whole it found the approaches implemented so far have had little or no effect on actual student learning, and in some cases have run counter to their intended purposes.



The results are likely to add fuel to ongoing debates across the country over how to fairly evaluate schools and teachers for student progress and whether to tie consequences for students and teachers to results from current forms of testing.

The study, released May 26, drew a mix of reactions.

"It's an antidote to what has been the accepted wisdom in this country, the belief that performance-based accountability and incentive systems are the answer to improving education," said Jon Baron, the president of the Washington-based Coalition for Evidence-Based Policy and the chairman of the National Board for Education Sciences, which advises the U.S. Department of Education's research arm. "That was basically accepted without evidence or support in NCLB and other government and private-sector efforts to increase performance," he said.

Eric A. Hanushek, an economics professor at Stanford University, said he was "stunned at how broad" the findings were. But he warned against using the committee's critique of test-based incentives to throw out accountability systems in education altogether.

"Some form of accountability is undoubtedly useful, but you have to be careful with how you structure accountability systems," Mr. Hanushek said. "What we've done to date hasn't been perfect; there are lots of obvious flaws in either results or program structure to date. As we go into the future, we should learn from our results."

Jim Bradshaw, a spokesman for the Education Department, said in an email: "This report confirms what

we already know—the accountability system in No Child Left Behind is broken and needs fixing this year. We need better assessments, college- and career-ready standards, and a more fair, focused, and flexible accountability system because children only get one shot at a world-class education.”

## Preventing Gaming

One critical flaw the study focused on was that test-based systems often use the same tests to gauge student progress and evaluate the system as a whole, with insufficient safeguards and monitoring to prevent educators or students from gaming the system to produce high scores disconnected from learning.

“Too often it’s taken for granted that the test being used for the incentive is itself the marker of progress, and what we’re trying to say here is you need an independent assessment of progress,” said Michael Hout, the sociology chairman at the University of California, Berkeley, and the chairman of the 17-member committee.

The panel, a who’s who of national experts in education law, economics, and social sciences, was launched in 2002 by the National Academies, a private, nonprofit quartet of institutions chartered by Congress to provide policy advice on science, technology, and health. Since its formation, the committee has been tracking the implementation and effectiveness of 15 test-based incentive programs, including:

- National school improvement programs under the No Child Left Behind Act and prior iterations of the Elementary and Secondary Education Act;
- Test-based systems of teacher incentive pay in Texas, Chicago, Nashville, Tenn., and elsewhere;
- High school exit exams such as those required by 28 states;
- Pay-for-scores programs for students in New York City and Coshocton, Ohio; and
- Experiments in teacher incentive pay in India and student and teacher test incentives in Israel and Kenya.

On the whole, the panel found the accountability programs often used assessments too narrow to accurately measure progress on program goals and used rewards or sanctions not directly tied to the people whose behavior the programs sought to change. Moreover, the programs often had inadequate checks in place to prevent manipulation of the system.

## Committee on Incentives and Test-Based Accountability

### Michael Hout (Chair)\*

Sociology Chairman  
University of California; Berkeley

### Dan Ariely

Professor of Psychology and Behavioral Economics  
Duke University; Durham, N.C.

### George P. Baker III

Professor of Business Administration  
Harvard Business School; Boston

### Henry Braun

Professor of Education and Public Policy;  
Director of the Center for the Student of Testing, Evaluation, and Educational Policy  
Boston College; Chestnut Hill, Mass.

### Anthony S. Bryk (until 2008)

President  
Carnegie Foundation for the Advancement of Teaching; Stanford, Calif.

### Edward L. Deci

Professor of Psychology and Social Sciences;  
Director of the Human Motivation Program  
University of Rochester; Rochester, N.Y.

### Christopher Edley Jr.

Professor and Dean of Law  
University of California; Berkeley

### Geno J. Flores

Former Chief Deputy, Superintendent of Public Instruction  
California Department of Education

“It’s not that there’s no information in the objective performance measures, but they are imperfect, and including the subjective performance measures is also very important,” said Kevin Lang, an economics professor at Boston University. “Incentives can be powerful, but not necessarily in the way you would like them to be.”

As a result, educators facing accountability sanctions tend to focus on actions that improve test scores, such as teaching test-taking strategies or drilling students closest to meeting proficiency cutoffs, rather than improving learning. Such a response undercuts the tests’ validity, the report says.

As an example, the report points to New York’s requirement that all high school seniors pass the state regents’ exam before graduating from high school. The policy led to more students passing the tests, but scores on the lower-stakes National Assessment of Educational Progress, which was testing the same subjects, didn’t budge during the same time period.

“It’s human nature: Give me a number, I’ll hit it,” Mr. Hout said. “Consequently, something that was a really good indicator before there were incentives on it ... becomes useless because people are messing with it.”

In fact, the study found that, rather than leading to higher academic achievement, high school exit exams so far have decreased graduation rates nationwide by an average of 2 percentage points.

The study found a growing heap of evidence that schools and districts have tinkered with how and when students take exit exams as well as other high-stakes tests in order to boost scores on paper for students who do not know the material—or to prevent those students from taking the tests at all.

### **AYP and Academics**

For similar reasons, school-based accountability mechanisms under the NCLB law have generated minimal improvement in academic learning, the study concludes. When the systems are evaluated—not using the high-stakes tests subject to inflation, but using instead outside tests, such as NAEP—student-achievement gains dwindle to about .08 of a standard deviation on average, mostly clustered in elementary-grade mathematics.

For perspective, an intervention considered to have a small effect size is usually about 0.1 of a standard deviation; a 2010 federal study of reading-comprehension programs found a moderately successful

### **Carolyn J. Heinrich**

Professor and Director of Public Affairs;  
Affiliated Professor of Economics  
University of Wisconsin-Madison

### **Paul T. Hill**

Research Professor; Director of the Center  
on Reinventing Public Education  
University of Washington Bothell

### **Thomas J. Kane\*\***

Professor of Education and Economics;  
Director of the Center for Education Policy  
Research  
Harvard University; Cambridge, Mass.

### **Daniel M. Koretz**

Professor of Education  
Harvard University; Cambridge, Mass.

### **Kevin Lang**

Professor of Economics  
Boston University; Boston

### **Susanna Loeb**

Professor of Education  
Stanford University; Stanford, Calif.

### **Michael Lovaglia**

Professor of Sociology; Director of the  
Center for the Study of Group Processes  
University of Iowa; Iowa City

### **Lorrie A. Shepard**

Dean and Professor of Education  
University of Colorado at Boulder

### **Brian M. Stecher**

Associate Director for Education  
Rand Corp.; Santa Monica, Calif.

\* Member, National Academy of Sciences

\*\* Was not able to participate in the final committee deliberations due to scheduling conflict.

SOURCE: National Academies

program had an effect size of .22 of a standard deviation.

Moreover, “as disappointing as a .08 standard deviation might be, that’s bigger than any effect we saw for incentives on individual students,” Mr. Hout said, noting that NCLB accountability measures school performance, not that of individual students.

Mr. Baron of the Coalition for Evidence-Based Policy said he was impressed by the quality of the panel’s research review, but unsurprised at the minimal results for various incentive programs.

Incorporating diverse types of studies—as the panel did—typically reduces the overall effects found for them, he noted.

“One of the contributions that this makes,” he said of the study, “is that it shows that looking across all these different studies with different methodologies and populations, some in different countries, there are very minimal effects in many cases, and in a few cases larger effects. It makes the argument that details matter.”

Committee members see hopeful signs in the 2008 federal requirement that state NAEP scores be used as an outside check on achievement results reported by districts and states, as well as the broader political push to incorporate more diverse measures of student achievement in the version of the ESEA that will revise the No Child Left Behind edition.

“It’s a message to all of us to slow down and think this through,” Jack Jennings, the president of the Center on Education Policy, in Washington, said of the findings. “We put all this weight on these tests that just weren’t designed for these things.”

He said the study is likely to focus lawmakers’ attention on the nearly \$400 million Race to the Top assessment grants, in which state consortia are developing testing systems to go along with the new common-core state standards. “There’s a lot riding on how these consortia do,” Mr. Jennings said.

Vol. 30, Issue 33, Pages 1,14

#### RELATED BLOG



[Visit this blog.](#)