

# High-Stakes Testing

## *The* **BALANCED VIEW:**

**Research-based  
information on  
timely topics**

**Volume 7, Number 1  
April 2003**

**WESTCHESTER  
INSTITUTE FOR  
HUMAN SERVICES  
RESEARCH**

**7-11 South Broadway  
White Plains, NY 10601  
(914) 682-1969  
FAX: (914) 682-1760  
e-mail:  
info@westchesterinst.org**

### *What is high-stakes testing?*

No issue in American education is more controversial than testing. Some people view testing as the linchpin of successful educational reform, while others see it as a threat to quality teaching and learning. Whatever the perspective, one thing is clear: more is riding on test performance today than ever before. One strategy, “*high-stakes*” testing, has come to dominate the educational landscape. Centrally featured in the accountability systems of many states and school districts, high-stakes testing involves rewarding or sanctioning schools, educators, and students on the basis of test results. It hinges on the motivational power of incentives to impel individuals to improve. Examples of “high-stakes” associated with testing include the following:

- ❖ *At the school level...*
  - distributing financial rewards to high-performing or improving schools
  - closing or restructuring low-performing schools
- ❖ *At the educator level...*
  - giving teachers or principals bonuses for high classroom or school test scores
  - replacing/removing teachers or principals for low classroom or school test scores
- ❖ *At the student level...*
  - promoting students to the next grade based on test results
  - granting students a high school diploma based on test results

- providing special diplomas or scholarships for college tuition to high-performing students
- giving students from failing schools the option to enroll elsewhere (e.g., through private school vouchers)

Given the significance of these and other consequences attached to test performance, there is a need for information about the effectiveness of high-stakes testing policies. This issue of the *Balanced View* summarizes what we know from the latest research.

### *Why is high-stakes testing important?*

Nearly 30 states now rate schools primarily or solely on the basis of student test scores. A similar number explicitly link student promotion or graduation to performance on state or district tests. In addition, more than half the states identify low-performing schools and have the power to close, take over, or reconstitute these schools; many also have the authority to replace school staff based on poor student performance. At the other end of the spectrum, a third of the states reward high-performing schools with money, and many give financial bonuses or pay raises to teachers and administrators for exemplary student performance. More states are expected to adopt such policies as a result of the *No Child Left Behind Act of 2001* (NCLB). NCLB requires all states to test students in grades 3-8 annually and

use the results to judge the performance of schools. Schools that fail to demonstrate sufficient improvement will be subject to serious sanctions.

These high-stakes policies dwarf all other educational change strategies in scope, and are fast becoming the end-all of state accountability reforms. With such widespread endorsement, it is essential that educators, the public, and policymakers critically examine the arguments for and against high-stakes testing as well as the evidence on its effects.

### ***What do people say about high-stakes testing?***

High-stakes testing is a decidedly contentious subject, with most voices either strongly for or strongly against it. Supporters say that high-stakes tests can bring greater coherence to the education system by clarifying student performance expectations. They claim that high-stakes tests provide a clear picture of what students need to learn to be successful and serve as a lever for holding all students and educators accountable to challenging standards.

Supporters also argue that high-stakes tests could be a significant factor in closing the achievement gap. By highlighting test score differences between rich and poor districts and among racial/ethnic groups, these tests can force educators to ensure that no group is “left behind.” This is especially so, say the advo-

cates, if rewards and sanctions are attached to the results. For without serious testing consequences, there would be little incentive for

a) test designs that exaggerate student achievement, b) “teaching to the test,” which may improve test-specific skills but also displace

**Supporters say that high-stakes tests produce better student achievement and narrow the achievement gap.**

**Critics counter that high-stakes tests do not promote real learning and result in unintended, negative consequences.**

schools to raise their expectations for failing children and address the achievement gap.

Finally, proponents of high-stakes tests argue that

- ❖ students work harder and learn more when they take high-stakes tests, and
- ❖ teachers are more motivated to focus their instruction on producing improved student achievement when tests have consequences.

At the same time, defenders acknowledge that these benefits are by no means assured. Not all teachers and students are equally motivated by high-stakes tests. And not all high-stakes tests are of equal quality. Some tests, for example, fail to measure the breadth and depth of state learning standards and end up promoting “drill and kill” teaching. But the best tests, say proponents, assess a complex array of content and skills, and, thus, encourage ambitious teaching that challenges all students.

Opponents, however, claim that high-stakes tests create adverse incentives that do not promote learning. They say that pressure to produce gains in test scores has led to

other skills important to real learning, and, c) a narrowing of the curriculum, where increasing emphasis is placed on the content sampled by the test, at the expense of deeper content and other valuable subjects, e.g., art, music, language.

Opponents further contend that high-stakes tests

- ❖ increase student grade retention and failure rates,
- ❖ result in higher dropout rates,
- ❖ are unfair to minorities, and
- ❖ lead to inappropriate labeling that can stigmatize children.

Some high-stakes testing critics even question the integrity of those administering the tests, suggesting that much of the claimed rise in student achievement is due to cheating. Lastly, detractors argue that good teaching and learning can occur in the absence of high-stakes testing; they cite evidence on the benefits of alternative strategies such as better-paid and trained teachers, ongoing professional development, and continued attention to challenging standards and curriculum.

## What does research say?

For a subject as contentious as high-stakes testing, there is surprisingly little empirical evidence on its effects. Much of the existing research is theoretical, subjective, or limited to a particular state or school district. Moreover, because state and/or district testing programs vary widely, it is difficult to draw firm conclusions from the results. Still, there are some recent studies on high-stakes testing that address questions of impact. We summarize the salient findings below.

**Stanford economists found that a two-step increase on their high-stakes accountability scale raised the achievement of minority students by 1 to 1.75 standard deviations. By way of comparison, the Tennessee STAR class-size experiment raised student achievement by 0.21 standard deviations.**

## Impact on Achievement

By and large, the recent evidence suggests that high-stakes testing produces higher student achievement. This is especially true in states that strongly punish and reward schools on the basis of student performance. For example, considerable achievement gains have been found in the following states: Alabama, California, Florida, Kentucky, Maryland, New Jersey, New Mexico, New York, North Carolina, and Texas—states that attach serious consequences to test results for schools *and* students. These positive effects have been demonstrated both via the

analysis of state tests and by drawing on data from the National Assessment of Educational Progress (NAEP), the nation's report card. The latter method is important because it addresses the issue of whether or not gains on high-stakes tests translate to other standardized tests—i.e., whether real learning is taking place.

A January 2003 report by Stanford University economists provides an understanding of the value added by high-stakes practices. These researchers found that a two-step increase on their *accountability index*<sup>1</sup> corresponded to a NAEP

achievement gain that exceeded .5 standard deviations for white 8<sup>th</sup> grade students.<sup>2</sup> Gains for black and Hispanic 8<sup>th</sup> graders were larger—1 to 1.75 standard deviations. At the 4<sup>th</sup> grade level, NAEP gains also were higher in states with more stringent accountability policies, although they were not as striking as the 8<sup>th</sup> grade results. Again, the rate of improvement of minority students in the “high-

<sup>1</sup>A 0 to 5 rating scale capturing the degree of state pressure on schools and students to improve achievement.

<sup>2</sup>Achievement gains of .25 standard deviations (SDs) or greater are generally considered educationally important. Gains ranging from .25 to .30 SDs are considered “small.” An increase of .30 to .40 SDs is considered “moderate.” An effect of .5 SDs or above is considered “large.”

stakes” states outpaced that of white students, signifying a closing of the achievement gap. To put these overall findings into perspective, it is useful to compare them to the widely-cited Tennessee STAR class size experiment; this popular reform raised student achievement by .21 standard deviations.

The Stanford University findings were in stark contrast to those reported a month earlier by Arizona State University (ASU) researchers. The ASU study, which received broad coverage in the popular and academic press, found that high-stakes testing produced few transferable academic gains and was a “failed policy initiative.” These conclusions were based on an analysis of NAEP, SAT, and other test data gathered from “high-stakes” states. A number of experts, however, have criticized the ASU study on methodological grounds, primarily for disregarding the magnitude of changes in test scores and for only examining states labeled as “high-stakes” rather than comparing states with strong and weak accountability systems.

## Impact on Student Retention and Dropout Rates

The research in this area is mixed. Some studies have shown increased retention and decreased graduation rates after the introduction of high-stakes tests. Others have not found this to be the case. Stanford University researchers uncovered little evidence that high-stakes testing increased student retention or adversely affected high school completion rates. On the other hand, the study found no evidence that high-stakes *improved* student retention or graduation rates. This, the researchers con-

cluded, was not “good news” for it suggests that high-stakes testing may not be changing the factors that influence educational attainment and progression through high school.

### High-stakes testing may not be changing the factors that influence educational attainment and progression through high school.

#### Impact on Curriculum and Instruction

The accumulated literature suggests that high-stakes testing has a powerful influence on curricular content. Specifically, as stakes increase, teachers spend more time covering tested areas and less time in non-tested areas such as fine arts, music, or foreign languages. The research is less clear on instructional impact. Several studies have shown that high-stakes testing encourages teachers to use methods that conform to the content of the test, which in itself is not a problem. The danger comes when the test measures only simple knowledge and skills, leaving complex concepts and reasoning underrepresented. Thus, “teaching to the test” can, as certain studies have reported, result in superficial, rote learning. It can also, as other studies have shown, result in higher-level thinking and problem-solving skills.

#### Costs

Studies on the costs of high-stakes testing are rare. One recent report from Harvard Uni-

versity concludes that current state accountability programs cost between \$5 and \$35 per pupil each year.<sup>3</sup> California and Texas, states with relatively comprehensive testing systems, for example, spend \$20 per pupil per year on their pro-

grams. In comparison, other educational reforms such as reduced class size can cost well-over \$500 per pupil. The costs of the Tennessee STAR class size reduction program, in fact, averaged \$3,500 per student. Thus, whatever the impact of high-stakes testing—be it positive or negative—it comes at a relatively low monetary cost.

#### Summary and Implications

Key findings from this research review can be summarized as follows:

- ❖ Student performance generally improves after high-stakes accountability reforms are introduced in states and districts.
- ❖ High-stakes testing can help to narrow the achievement gap.
- ❖ Effects of high-stakes testing on student retention or high school graduation rates are unclear. To date, there is no solid evidence that high-stakes testing either improves or worsens graduation or student retention rates.
- ❖ High-stakes testing leads to an increased emphasis on tested

content, often at the expense of non-tested subjects.

- ❖ The impact of high-stakes testing on instruction depends on the format of the tests. Tests that measure complex concepts and extended reasoning encourage stimulating instruction, and vice versa.
- ❖ The impact of high-stakes tests comes at a relatively low cost.

A reasonable conclusion from this evidence is that high-stakes testing policies are making a difference in the area of achievement, especially for minority students. At the same time, they seem to be leading to a narrowing of the curriculum, and may have no appreciable impact on student progression through high school. These conclusions, however, must be considered cautiously for much remains to be learned about the impact of such a large-scale reform. In the meantime, policymakers and educators will have to determine if results such as these are the desired ones. To the extent that undesired effects are occurring, testing programs may need to be modified. Two resources for understanding how best to use high-stakes testing are identified below:

American Psychological Association:

<http://www.apa.org/pubinfo/testing.html>

American Educational Research Association:

<http://www.aera.net/about/policy/stakes.htm>

**The *Balanced View* welcomes your comments. Primary references used are available upon request.**

<sup>3</sup> These figures include the cost of the assessment, itself, as well as the cost of running an office of accountability.