

They were told that science was a stepping stone to all sorts of learning and how much students loved it.

But I saw very little science in the third grade at Tyler Heights. The kits in Johnson's room would be opened to roll marbles one time early in the year, and later to make goo and sculpt a landform and to compare seeds and pebbles in a petri dish. These were only a tiny fraction of the experiments inside, and at any rate, they were presented in class severely abridged—no hypotheses, no data. Mostly students read from the textbook and did worksheets. The only full pass through the scientific method was made after the MSA, in the days spent preparing for the science fair.

"I'm a realist," McKnight had told the teachers. "What gets taught is what gets tested." The rest—even if it is part of the state standards—gets left behind. When it came to the accountability movement, McKnight epitomized the ambivalence of most educators I've met: she was supportive of standards and testing in theory, but painfully aware of the unintended consequences. She was passionate about the

subject she used to teach—social studies, and particularly geography—but when it came down to it, social studies fared no better than science.

Tyler Heights' third-graders got only the most cursory introduction to economics and Native Americans, and much of the curriculum was skipped altogether. The students were geographically ignorant. Approaching the Naval Academy after a three-mile bus ride, several shouted, "Look, it's New York!" The third-graders had heard Africa mentioned a lot but were not sure if it was a city, country, or state. (They never suggested "continent.") At the end of the year, the children in Johnson's class were asked to name all the states they could. Cyrus knew the most: three. He couldn't name any countries, though, and when asked about cities, he thrust his finger in the air triumphantly. "Howard County!"

McKnight had asked teachers to give students passages on social studies and science topics for supplemental reading lessons in preparation for the MSA. But the passages the third-graders read touched

on random knowledge—Billie Holiday's alcoholism, female Arctic explorers—and breezed by quickly. They were hard to understand on the fly when the children had had such little exposure, at school and at home, to history, culture, and the natural world.<sup>†</sup>

\* \* \*

At a conference on assessment, a reading specialist from the Maryland Department of Education told teachers and principals desperate to unlock the secrets of the MSA that BCRs are not tests of writing skills at all, but of reading. "I'm not saying kids shouldn't write well-developed paragraphs," she told the standing-room-only crowd. "But that's not what we're worried about on this test."

"You could bullet it, list key phrases, and you could get the same number of points as someone who wrote a well-crafted answer," McKnight said. The formula is a helpful scaffold, she said, but "if the only

<sup>†</sup>For the relationship between background knowledge and reading skills, see E. D. Hirsch, Jr., *The Knowledge Deficit: Closing the Shocking Education Gap for American Children* (New York: Houghton Mifflin, 2006).

## Tyler Heights Is Not Alone

### *Score Inflation Is Common in Education—and Other Fields*

BY DANIEL KORETZ

Every year, newspaper articles and news releases from education departments around the nation tell us that test scores are up again, often dramatically. Usually, there are some grades or districts that have not made substantial gains, and the gaps in performance between poor and rich, and majority and minority, often fail to budge. Nevertheless, the main story line is usually positive: performance is getting better, and rapidly.

Unfortunately, this good news is often more apparent than real. Scores on the tests used for accountability have become inflated, badly overstating real gains in student performance. Some of the

reported gains are entirely illusory, and others are real but grossly exaggerated. The seriousness of this problem is hard to overstate. When scores are inflated, many of the most important conclusions people base on them will be wrong, and students—and sometimes teachers—will suffer as a result.

This is the dirty secret of high-stakes testing. You may see occasional references to this problem in newspapers, but for the most part, news reports and announcements of scores by states and school districts accept increases in scores at face value.

When I and others who work on this issue point it out, the reactions often range from disbelief to anger. So perhaps it is best to start on less controversial ground. We see something akin to score inflation in many other fields as well. It is so common, in fact, that it has the name Campbell's law in social sciences: "The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures

and the more apt it will be to distort and corrupt the social processes it is intended to monitor."<sup>1</sup> One can find examples of Campbell's law in the media from time to time that provide a hint of how score inflation arises.

The most disturbing example of Campbell's law that I have encountered was reported by the *New York Times* in 2005. The School of Medicine and Dentistry at the University of Rochester had surveyed cardiologists around the state. As the *Times* reported, "An overwhelming majority of cardiologists in New York say that, in certain instances, they do not operate on patients who might benefit from heart surgery, because they are worried about hurting their rankings on physician scorecards issued by the state."<sup>2</sup> Fully 83 percent of respondents said that the reporting of mortality rates had this effect, and 79 percent admitted that "the knowledge that mortality statistics would be made public" had affected their own decisions about whether to perform surgery.\*

*Daniel Koretz is the Henry Lee Shattuck Professor of Education at Harvard University's Graduate School of Education and a member of the National Academy of Education. This sidebar is excerpted with permission from Measuring Up: What Educational Testing Really Tells Us, Harvard University Press, copyright © 2008 by the President and Fellows of Harvard College.*

thing you're teaching is BCRs, your kids are not learning to write."

The third-graders at Tyler Heights, then, did not learn to write. They learned, thanks to a timer broadcast on the overhead projector, to fill in the box of eight lines in seven to nine minutes. They learned to "proof and polish" with a special purple pen, and whisper their paragraphs to themselves through C-shaped sections of PVC pipe held to their ears—what they called "whisper phoning," a strategy for detecting if your answer makes sense. They learned to adhere to the BATS formula in BCRs like the one Johnson led her students through one day:

*Damon and Pythias is a play because it has the elements of a play. Some elements of a play are that plays have stage directions. Also, there is a narrator. This play also has a lot of characters. So I know this play has all the features it needs.*

The BCRs tended to repeat themselves, so the children worked on a limited range of questions teachers knew would be on

the county benchmark tests and suspected would be on the MSA. The third-graders answered again and again what traits described the main character of a story. They wrote the "I know this is a play because" BCR about 10 times but never got to act out a play. They wrote "I know this is a fairy tale because" and "I know this is a fable because" but never tried their hand at creating either. About a fake brochure they wrote, "The text features that make this easy for a third-grader to understand are italics, numbering, and underline." But they never made their own brochures with their own text features; the only things they underlined were hundred-dollar words. They wrote "I know this is a poem because it has rhyme, rhythm, and stanzas" about 50 times, Johnson estimated, but they only wrote three poems.

The Tyler Heights teachers knew that the BCR focus was a problem but were either unwilling or unable to veer from the program—they felt they were not allowed. One day in the teachers' lounge, two for-

mer teachers who were now an aide and a mentor reminisced about the days when third-graders read novels and did chemistry experiments and worked in groups to design versions of the 13 colonies and did writing, real writing. A resource teacher who was an active part of the school's laser-sharp focus over the last few years began to question her own role. She listened to the veterans and added her two cents. "While our scores were really good last year, can I tell you our kids are any smarter? I don't know."

\* \* \*

Tyler Heights was not explicitly ordered to de-emphasize topics that are not tested; then again, nobody from the school district, and nobody who lauded the school for its scores, bothered to make sure the whole curriculum was taught. On the last day of MSA testing, McKnight said to me, "MSA, that's just the bottom of what kids should know. It's not like we were calling them brilliant. We're still shooting for the basement. We celebrate the bottom right now. I pray we don't have to keep celebrating the bottom." □



So it should not be surprising that when the heat is turned up, educators—and students—will sometimes behave in ways that inflate test scores. Actually, it would be quite remarkable, given how pervasive the problem is in other fields, if none of them did.

Advocates of current test-based accountability systems often counter by arguing, "So what if the gains are distorted? What matters is that students learn more, and if we get that, we can live with some distortion." Hypothetically, yes, we could live with it if we knew that students were in fact learning more, and if the distortions were small enough that they did not seriously mislead people and cause them to make incorrect decisions. But in fact, we usually cannot distinguish between real and bogus gains. Because so

many people assume that if scores are increasing we can trust that kids are learning more, there is a disturbing lack of good evaluations of these systems, even after more than three decades of high-stakes testing. What we do know is that score inflation can be enormous, more than large enough to seriously mislead people.

As a result, we need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. They are likely instead to produce substantial inflation of scores and a variety of undesirable changes in instruction, such as an excessive focus on old tests, an inappropriate narrowing of

instruction, and a reliance on teaching test-taking tricks.

I strongly support the goal of improved accountability in public education. I saw the need for it when I was an elementary school and junior high school teacher many years ago. I certainly saw it as the parent of two children in school. Nothing in more than a quarter century of education research has led me to change my mind on this point. And it seems clear that student achievement must be one of the most important things for which educators and school systems should be accountable. However, we need an effective system of accountability, one that maximizes real gains, and minimizes bogus gains and other negative side effects.

In all, educational testing is much like a powerful medication. If used carefully, it can be immensely informative, and it can be a very powerful tool for changing education for the better. Used indiscriminately, it poses a risk of various and severe side effects. □

#### Endnotes

1. Donald T. Campbell, "Assessing the Impact of Planned Social Change," in *Social Research and Public Policies: The Dartmouth/OECD Conference*, ed. Gene M. Lyons (Hanover, NH: Public Affairs Center, Dartmouth College, 1975), 35.
2. Marc Santora, "Cardiologists Say Rankings Sway Choices on Surgery," *New York Times*, January 11, 2005.

\*These numbers may be off by a modest amount, but not by enough to make the results less appalling. Only 65 percent of the sampled surgeons responded to the survey, which is a marginally acceptable response rate. The risk is that surgeons who did not respond would have given different answers than those who did. But even if all 35 percent who did not respond would have replied to these questions in the negative—an extremely unlikely case—that would still leave more than half saying that publication of mortality measures led to surgeons' declining to do procedures that could have benefited patients.