HIGH-STAKES research

by MARGARET E. RAYMOND &
ERIC A. HANUSHEK

"Make-or-Break Exams Grow, But Big Study Doubts Value" intoned a front-page *New York Times* headline in December 2002. The article continued, "Rigorous testing that decides whether students graduate, teachers win bonuses, and schools are shuttered … does little to improve achievement and may actually worsen academic performance and dropout rates, according to the largest study ever on the issue." Thus a deeply flawed study was catapulted to national prominence. More important, its conclusions were opposite those found through rigorous scientific studies.

The report in question, authored by Arizona State University researchers Audrey Amrein and David Berliner, purported to examine student-performance trends on national exams in states where legislators have attached "high stakes" to test scores. High-stakes testing has become a lightning rod as more and more states adopt accountability measures in response to the mandates of the federal No Child Left Behind Act. While it is crucial to analyze and debate the wisdom of such poli-

The campaign against accountability has brought forth a tide of negative anecdotes and deeply flawed research.

Solid analysis reveals a brighter picture

ILLUSTRATION BY ANDREW JUDD/MASTERFILE

cies, the discussion must be informed by evidence of the highest quality. The controversial nature of high-stakes testing has led to the hurried release and dissemination of research that lacks scientific rigor, of which the Amrein and Berliner study is one of the more egregious examples.

This says much about the standards for research in education today. The situation is so contentious that in 2000 the National Research Council found it necessary to convene a panel to decide which scientific principles should apply to educational research—the kind of question that other fields of social science settled long ago. In the case at hand, Amrein and Berliner trumpet the fact that their report was reviewed by a panel of four scholars based at other schools of education, yet this should only be a source of greater concern. Sharing a paper with sympathetic colleagues is no substitute for a system of blind peer review—a bedrock principle of scientific research.

Amrein and Berliner's misleading reporting practices took on new importance when the media dutifully broadcast the results as they were written.

Here we closely examine Amrein and Berliner's underlying data and methodology. Our results are astonishing: if basic statistical techniques are applied to their data, it reverses nearly every one of their conclusions. Later we also present the results of separate research on accountability that we conducted for a June 2002 Federal Reserve Bank of Boston conference. Rigorous analysis reveals that accountability policies have had a positive impact on test scores during the past decade.

The Unscientific Method

Amrein and Berliner identified 28 states where test scores are used to determine various consequences, such as bonuses for teachers, the promotion of students, or allowing children to transfer out of a failing school. These stakes go beyond less controversial accountability measures such as publishing test scores in the newspaper. The states range from Georgia and Minnesota—where the only penalty is experienced by students who fail a high-school graduation exam—to North Carolina and Texas, where the authors found a total of six stakes each, stakes that affect both schools and students.

Once Amrein and Berliner identified the high-stakes states, they looked at changes in the average scores students earned on the National Assessment of Educational Progress (NAEP). Choosing this test as a basis for considering the impact of high-stakes tests on students in the 4th and 8th grades (ages 9 and 13, respectively) is a sensible idea, because the validity and reliability of NAEP, often called the "nation's report card," are well accepted. It is a test for which students cannot easily be prepped and, since the performance of individual school districts, schools, or students is not reported, there is little incentive to cheat or even to prepare for the test. It also provides a neutral standard for assessing the effects of state policies. But if the Arizona State team's decision to look at NAEP scores was correct, less can be said for their other analytical choices.

Amrein and Berliner's basic strategy was to look at how each high-stakes state's scores changed with the introduction of accountability and to compare this with the national trend. If the state's gains exceeded the national gains, they deemed that an increase in scores. If the state's gains trailed the national gains, they deemed that a decrease. But whenever the rate at which students were excluded from the NAEP because of a disability or lack of language proficiency moved in the same direction as that state's NAEP scores (in other words, an increase in test scores coupled with an increase in test exclusions), Amrein and Berliner declared the results contaminated and simply tossed out the state as inconclusive. (At least that is what they claimed to do; in fact, they applied the rule inconsistently.)

As a result, their conclusions are based on only a fraction of the high-stakes states. For instance, they recorded positive or negative results on the NAEP 4th-grade math test for just 12 of the 26 states with stakes for K–8 students (as noted earlier, two of the states, Georgia and Minnesota had only a high-school graduation exam and thus were not used for this analysis). Amrein and Berliner found that 4th-grade math scores increased at a slower rate than the national average in 8 of the 12 states, faster in just 4. Yet they write this up in a highly misleading fashion, claiming that "67 percent of the states posted overall decreases in NAEP

Rerunning the Amrein-Berliner Data (Table 1)

When the actual test scores in the states Audrey Amrein and David Berliner identified as "high stakes" are compared with those in states without accountability systems, the high-stakes states show much more improvement.

	Increase in NAEP 4th-grade math scores		Increase in NAEP 8th-grade math scores	
	1992-2000	1996-2000	1992-2000	1996-2000
High-stakes states	9.2	4.2	8.8	4.5
No accountability states	3.8	2.3	4.0	1.7
High-stakes advantage	**5.3 points***	**1.9 points***	**4.8 points***	**2.8 points***
High-stakes advantage after adjusting for changes in students excluded from NAEP	5.2 points*	2.3 points*	3.7 points*	2.5 points*

* statistically significant at the .05 level

SOURCE: Authors

math grade 4 performance as compared to the nation after high-stakes tests were implemented." Actually, Amrein and Berliner witnessed gains slower than the national average in just 8 of 26 high-stakes states, or 31 percent.

Amrein and Berliner's misleading reporting practices took on new importance when the media dutifully broadcast their results as they were written. Consider the article in *Education Week*, which reported, "Movement in elementary-school reading scores was evenly split—better than the national average in half the states, worse in the other half." In fact, Berliner and Amrein based their conclusions in 4th-grade reading on just ten states, five of which they recorded as gaining against the national average, five of which as losing. So less than a fifth of the high-stakes states saw decreases against the national average in reading, not "half." At the 8th-grade level in math, Amrein and Berliner were able to look at only eight states, five of which gained against the national average, three of which lost. Here, again, Amrein and Berliner wrongly reported this as "63 percent of the states posted increases in NAEP math grade 8 performance as compared to the nation after high-stakes tests were implemented."

All of this ignores the truly fatal flaw of Amrein and Berliner's methods: their point of comparison. If one wants to assess the effect of high-stakes testing, the obvious comparison is between states that adopted accountability systems and those that did not. Amrein and Berliner's decision instead to compare the gains in high-stakes states with the national average violates a basic principle of social-science research. The national gain on NAEP incorporates any gains in high-stakes states, so Amrein and Berliner's strategy is akin to a medical trial where the treatment group receives the full dose of a medication while the control group receives a half-dose. It would not be surprising to find that the full dose was not dramatically more effective. The real question is whether the full dose is more effective than no medication at all.

On Their Terms

Amrein and Berliner concluded, as announced in their press release, "High-stakes tests may inhibit the academic achievement of students, not foster their academic growth." Let's take a look at their evidence in more detail.

Before doing so, however, we need to be clear: we are not in any way endorsing Amrein and Berliner's analytical approach. We return below to discuss the results from a more scientific study of accountability. But using their approach in a systematic manner will at least reveal the degree to which their decisions about what information to include and to exclude distorted the facts and thereby confused the debate over accountability.

An initial problem with their analysis is that Amrein and Berliner disregarded the magnitude of any changes in test scores. By simply listing the results as "Increase," "Decrease," or "Unclear" (in cases where exclusion rates rose), Amrein and Berliner discarded rich information. They converted useful continuous data (test scores) into hollow binary data (test scores went up or down). In a purely hypothetical example, say six of the high-stakes states gained 20 percent, while the other 20 gained 2 percent each and the no-accountability states made no gains whatsoever—yielding a national average gain of 3 percent. Amrein and Berliner's approach would supposedly demonstrate the failure of accountability: just six states beat the national average, while 20 were below the

A Panoply of Mistakes (Table 2)

Test scores actually increased at a faster rate than in no-accountability states in almost all of the high-stakes states where Audrey Amrein and David Berliner (AB) claimed to find decreases in scores. In New Mexico, Oklahoma, and West Virginia, where AB found decreases, high-stakes testing was introduced too early to make a valid before-and-after comparison.

States where AB declared decreases in NAEP scores	Introduction of high-stakes testing (AB date)	Change in 4th-grade NAEP math scores between:		
		1992-1996	1996-2000	1992-2000
Kentucky	1994	4.9[b]	1.0	5.9[c]
Maryland	1993	3.4[b]	1.6	5.0[c]
Missouri	1993	2.5[c]	3.8[b]	6.3[c]
Nevada	1998	N/A	2.7[c]	N/A
New Mexico	1989[a]	0.5	0.0	0.6
New York	1999	4.2[b]	3.9[b]	8.1[b]
Oklahoma	1989[a]	N/A	N/A	4.7[c]
West Virginia	1989[a]	8.1[b]	1.5	9.6[b]

Notes:

N/A – NAEP data unavailable for this time period

a. No NAEP tests at or before introduction of high-stakes testing

b. Change in NAEP scores exceeds the average change in NAEP both for the nation and for states not adopting high-stakes testing

c. Change in NAEP scores exceeds the average change for states not adopting high-stakes testing

SOURCE: Authors

average. In fact, ignoring any complications from test exclusions, Amrein and Berliner would report this as something like, "Just 23 percent of states posted gains on NAEP higher than the national average after high stakes were introduced." The right approach is to compare the average gains of high-stakes states with those of no-accountability states.

When this is done, the analysis yields starkly different results than Amrein and Berliner report. Table 1 compares the math gains among 4th and 8th graders in the same way as Amrein and Berliner—by following different cohorts as they reach 4th or 8th grade in different years. In other words, they compared the 4th graders of 1996 with the 4th graders of 2000, two completely different cohorts of students. For each of the comparisons, data were available for 34–36 states, 18–20 of which were part of the high-stakes group, due to the varying participation of states in the NAEP testing program. For either the 1992–2000 period or the 1996–2000 period, the average gain in math among high-stakes states noticeably exceeded that of the no-accountability states. The differences in performance were statistically significant at conventional levels, meaning that we can be highly confident that they are not just chance occurrences. (By contrast, Amrein and Berliner did no significance testing whatsoever, neglecting one of the oldest and most basic tools of social-science research.)

Amrein and Berliner might object that we have included states where students were excluded from tests at higher rates after accountability reforms were introduced, possibly contaminating the results. Amrein and Berliner's solution was just to toss these states out, no matter how small the change in exclusion rate or how large the change in achievement. As Table 1 shows, we instead adjusted the achievement gains for observed changes in exclusion rates. And the results barely changed: high-stakes states still significantly outperformed no-accountability states across the board. In fact, the changes in test-participation rates were not statistically different in high-stakes states from those in other states, indicating that this was not even remotely as influential a factor as Amrein and Berliner declared it to be.

Scientific quality is determined not only by the overall methodology, but also by the care and precision of any measurements. To assess the latter, let's focus on the eight states where Amrein and Berliner concluded that 4th-grade math scores decreased following the introduction of high-stakes testing. Consider Table 2. Three of the eight states—New Mexico, Oklahoma, and West Virginia—adopted high-stakes testing during the 1980s. However, NAEP scores at the state level became available only during the 1990s. For these states, Amrein and Berliner lacked

Correctly applying Amrein and Berliner's underlying approach to all of the data on NAEP achievement reverses their conclusions.

the "before" data for their "before and after" analytical strategy, but went ahead and labeled their scores as "decreasing" anyway. The other five "decreasing" states all experienced greater gains than no-accountability states during the time that they introduced high-stakes testing; New York even beat the national average gain in every time period. And this is the group of states that Amrein and Berliner identify as being harmed by accountability! *Not a single one* provides evidence of harm following the introduction of high-stakes testing.

Even where before-and-after data were available, Amrein and Berliner did not always use the data from the NAEP tests immediately preceding and following the adoption of high stakes. In several cases, they apparently chose an interval that began after the state's accountability system came on-line—an "after-after" comparison. These procedures yielded results that reflected negatively on accountability, but they have no scientific justification. To see this, consider the table on p. 52 and try to think of a consistent rule that justifies Amrein and Berliner's decision to place both Maryland and Missouri in the "decreasing" category.

In short, Amrein and Berliner used scientifically inappropriate methods and applied them in an even shoddier manner. Simply taking Amrein and Berliner's approach and applying it correctly to all of the data on NAEP achievement reverses their conclusions. Again, these simple comparisons are not the best way to examine these questions, but the results of even these crude analyses confirm the findings from the more sophisticated approach we describe below: greater accountability is accompanied by improved student performance.

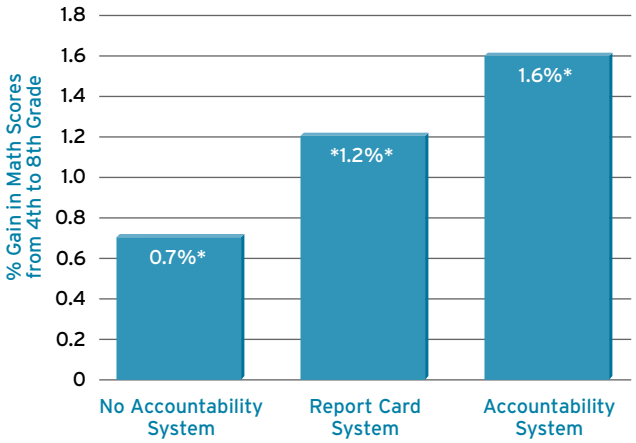
Amrein and Berliner also used trends on the SAT, the ACT, and Advanced Placement (AP) exams to assess the effectiveness of minimum-competency exams in the 18 states where students must pass such tests in order to graduate from high school. This comparison suffers from all the same problems as the NAEP comparison and more. For example, does anyone believe that nothing else has changed in North Carolina since the introduction of a graduation test in 1980? Amrein and Berliner's simplistic trend analysis attributes all subsequent changes in graduation rates and dropouts to the introduction of this high-stakes exam. Nonetheless, because these discussions are less directly related to the current state accountability debates and these data are more difficult to interpret than NAEP scores, we do not pursue them.

Results of Rigorous Analysis

Assessing the impact of state accountability systems is clearly complicated. In many states, these systems are quite young; in 1996, just ten states had active accountability systems. Moreover, states differ in many ways other than their accountability provisions—ways that can make it difficult to isolate the impact of high-stakes testing. They also change in different ways over time, adopting new accountability provisions and other legislation at different times and being influenced by shifting demographics at different rates. This does

Accountability Works (Figure 1)

States that reward or sanction schools for their academic performance made greater gains on the National Assessment of Educational Progress in math from 1996 to 2000.



*Statistically significant at the .05 level

SOURCE: Authors

States with high-stakes and even low-stakes accountability systems for schools performed significantly better on NAEP than states with no stakes at all.

not make gathering evidence about the effects of accountability impossible. It simply reinforces the need to apply stringent scientific methods to the analysis.

Here we report results from our own analysis of state accountability systems using NAEP data. These results were reviewed at a high-profile conference and were subject to a blind peer review for publication in a Brookings Institution volume, *No Child Left Behind? The Politics and Practice of Accountability*, which is slated for release this fall.

NAEP tested 4th graders in mathematics in 1992 and 1996 and 8th graders four years after each of these assessments, in 1996 and 2000. As noted earlier, whereas Amrein and Berliner simply compared the test scores of 4th graders in one year with those of a different set of 4th graders four years later, we measured students' growth in achievement between the 4th and 8th grades. In other words, we compared 4th graders' math achievement in 1996 with their performance four years later, when they were 8th graders. The same exact students were not tested in each grade, but the two samples are at least representative of the same cohort of students. We also adjusted the data to account for changes in state spending on education and for parents' educational levels, which provides controls for simultaneous changes in state policies or differences in demographics that might confound the analysis of how accountability systems influenced student achievement. Amrein and Berliner used no statistical controls at all.

Our analysis focuses on state testing and accountability systems that impose consequences on schools rather than on students. These are the most relevant policies for evaluating the potential impact of the No Child Left Behind Act. Our statistical analysis includes all states that have relevant NAEP data, and we explicitly allow for the timing of states' introduction of their accountability systems.

Figure 1 summarizes our findings in mathematics. The typical student progressing from grade 4 in 1996 to grade 8 in 2000 in a state with a consequential accountability system could expect to see a 1.6 percent increase in his NAEP proficiency score (calibrated to the appropriate learning standards for each grade). By contrast, the typical

student in a state with no accountability system could expect to experience only a 0.7 percent gain in mathematics proficiency, a statistically significant difference. Students in states with "report card" systems, where scores are publicly reported but no consequences are attached to performance, fell in the middle: they could expect to gain 1.2 percent in achievement between grades 4 and 8, over and above what they would normally learn from grade to grade. In short, states with high-stakes and even low-stakes systems for schools performed significantly better on NAEP than states with no stakes at all.

We are not the only ones reporting positive effects of accountability. In a forthcoming paper, Stanford University economists Martin Carnoy and Susanna Loeb conducted a similar analysis but expanded it to include testing policies

that impose high stakes on students. They found that NAEP performance increased in states with high-stakes systems compared with states that had not yet attached consequences to schools' test scores. Carnoy and Loeb also investigated the impact of accountability on student retention and high-school graduation rates and demonstrated that there is no discernible negative effect on either outcome.

Other rigorous studies have been carried out of accountability systems within states and school districts. As opposed to the Amrein-Berliner study, they have been vetted at scientific conferences and are being peer reviewed according to normal scientific practice. The Brookings Institution volume is one example. The accumulated literature generally supports two conclusions. First, student performance on the available measures, usually state tests, improves after accountability reforms are introduced. Second, other short-run changes—such as students' being excluded from taking the tests at greater rates, or explicit cheating—are observed. In other words, some unintended consequences often tend to accompany the introduction of accountability, although there is little evidence that these influences continue over time.

Schools may exclude low-performing students from taking the test in an attempt to "game" the system—to increase their performance artificially by removing scores that bring down their averages. We looked at differences among the states in terms of their placement rates into special education—often one way to exclude students from state tests—and at whether these differences were related to the introduction of state accountability systems. From 1995 to 2000, the time when many state accountability systems were coming on-line, we found no evidence that special-education placement increased in reaction to the introduction of accountability. Special-education placements did increase nationally, just not in any systematic way suggestive of a relationship to state accountability.

No Accountability for Research

That a study of such dubious scientific quality could make the front page of the nation's most respected newspaper is disturbing, but perhaps not so unusual. In the contentious environment of K–12 education, the media too often gives attention to findings that are relevant to policy regardless of their scientific merit. This discussion shows that education studies vary so much in their scientific rigor that one cannot just review them based on press releases and the sensationalism of the reported results.

Reporters need not be experts in statistical analysis any more than they must be fully versed in biochemistry or investment-banking regulations. But when a report is commissioned by an organization like the Great Lakes Center for Education Research and Practice, a Midwestern group sponsored by six state affiliates of the National Education Association, it would seem to call for a reasonable dose of skepticism. Why not bring in some outside expertise to review such a report before heralding its arrival? There will definitely be further opportunities for review. After all, the Arizona State shop promises that this is just the first of many annual reports on the impact of high-stakes testing.

The media is not alone. Resources at the state and federal levels must be committed to evaluating the quality of research and disseminating evidence of effective practices to schools and the public. The No Child Left Behind Act's emphasis on research-based practices, the creation of the federal Institute of Education Sciences, and efforts such as the What Works Clearinghouse, which will review and disseminate research findings, are important developments in this regard. State policymakers must also devote resources to evaluating their programs and synthesizing available research. Identifying effective reforms using rigorous evaluative techniques is a crucial task, especially since improving the education system is likely to have a greater economic impact than any of the medical breakthroughs of the past decade.

We also do not mean to suggest that the book has been closed on accountability. It appears that high-stakes states performed better than no-accountability states during the 1990s, but there is still much to be learned. For instance, there is uncertainty about the best way to translate test scores into overall school ratings. Also, states have yet to design accountability systems that directly link test-score performance to appropriate incentives. The vast majority of state accountability systems simply report the average scores for each school, sometimes disaggregating by racial and ethnic groups. However, average scores are highly dependent on socioeconomic factors outside the control of schools. States—and researchers—must become adept at discerning the components that make up the scores and how they can be influenced by high-stakes regimes. Measuring the gains that students make over time would provide a better measure of school performance and serve as a proper basis for reward or sanction, but such value-added techniques need some work before they can serve as reliable performance measures. There are other issues as well. Nonetheless, the evidence points in the direction of refining accountability systems rather than scrapping them altogether.

–Margaret E. Raymond is the director of CREDO, an education policy research group at the Hoover Institution. Eric A. Hanushek is a senior fellow at the Hoover Institution.