Shopping for Evidence Against School Accountability

Margaret E. Raymond
Eric A. Hanushek
Stanford University

About the Authors

Margaret E. Raymond is director of CREDO at the Hoover Institution of Stanford University. CREDO provides impartial evaluation of educational programs and policies. She has published independent evaluations of Teach for America, charter schools, and California's accountability system. She can be contacted at macke@stanford.edu.

Eric A. Hanushek is the Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University, chair of the executive committee for the Texas Schools Project, and a research fellow of the National Bureau of Economic Research. He has written extensively about the economics and finance of schools. He can be contacted at hanushek@stanford.edu.

Shopping for Evidence Against School Accountability

Margaret E. Raymond
Eric A. Hanushek
Stanford University

Accountability has been a central feature of educational policy in a number of states since the 1990s. In part because of the perceived success of accountability in the states where it was initially tried, federal law introduced mandatory reporting and accountability through the No Child Left Behind Act of 2001. Yet not everybody is happy with school accountability. Its opponents continue to aggressively search for evidence that testing and accountability do not work—or, better, that they are actually harmful. The hope of the anti-accountability forces is that they can stop testing before it is fully in place and before rollbacks would be impossible.

The window of opportunity to cripple or stop testing is narrowing over time, so it is not surprising that hasty reports based on biased research should appear. Nor is it surprising that these reports are given attention by parties who are unschooled in the requirements of good research. Perhaps we could disregard these events if the policies themselves were unimportant or if public exposure to poor quality studies had no effect on the ultimate decisions about them. But that is not the case. Since testing and accountability represent the cornerstone of current school reform efforts, it is essential that we apply rigorous standards of evidence and of scientific method to the analysis of accountability policy. The impact of testing and accountability is perhaps the most important issue facing school policymakers today. Even though accountability, by itself, does not say anything about how to organize an effective school, measures of school performance provide a standardized construction of information needed to forge through the bewildering array of "answers" to the question of how to improve our schools. While it is certainly reasonable to question the effectiveness of particular accountability systems and the policy of accountability in general, little thought has been given to the scientific standards of evidence that ought to apply to research and evaluation aimed at informing or influencing the policy process in this important area.

Assessing the impact of state accountability is clearly difficult. Policies have been in place for a limited amount of time. All states but one have adopted a system in one form or another. Not all accountability systems are the same. When put in place, they apply to all schools within entire states, limiting relevant variation to differences across states. This means that we have lost forever the chance to test whether accountability systems are superior to what states had before. Finally, accountability systems are just one of many ways in which states tend to differ. These factors do not imply that gathering evidence about the effects of

accountability is impossible. They simply reinforce the need to apply strict scientific methods to ensure that uncertainty is reduced as far as possible.

Bad news about accountability gets an undue amount of media coverage. First, the anti-accountability forces trumpet any possible scrap of data that might be portrayed as generalizable evidence against routine testing and accountability. Second, researchers reinforce this by their popular search for unintended consequences of government actions. Finally, the press, looking for both controversy and balance in reporting, tends to cite any study—no matter what its scientific quality—to show the evenhandedness of its reporting.

What do we know to date? The existing evidence on state accountability systems indicates that their use leads to improvement of student achievement. States that introduced accountability systems during the 1990s tended to show more rapid achievement gains when compared to states that did not introduce such measures. Along with general improvement, there also appear to be instances of unintended consequences—such as increased special education placement or outright cheating—at the time of introduction, but there is no evidence that this continues over time. Looking across states, we also know that attaching stakes to performance on tests yields better performance. Though still preliminary, these findings rest on rigorous analytic techniques, providing policymakers the most reliable evidence yet available.

The existing evidence on state accountability systems indicates that their use leads to improvement of student achievement.

What do we not know to date? Plenty. We do not know which general designs of accountability systems work best, or even the best underlying content standards for achievement. Nor do we know the optimal way to attach rewards and punishments to performance. Who should be judged by what scores? These are things that will take time to discover, but there is no way to get from here to there without a systematic approach to future policy enhancements and continued rigorous evaluation of their effects.

Evidence About Existing Accountability Systems

Over the past decade, states have devised diverse accountability systems that differ by choice of test, grades monitored, subjects tested, and performance requirements. Direct comparison of state against state based on state accountability system information is therefore problematic; a common but independent standard of comparison is needed. One source of information on performance, however, offers some possibility for analysis. The National Assessment of Educational Progress (NAEP), the "Nation's Report Card," provided performance information for states during the 1990s. While not designed as a national test, these examinations provide a highly respected and consistent tracking of student performance across grades and time. Since scores are not reported for individuals or schools, there is no incentive to prep for them or to cheat on them. We have used these performance measures to assess the impacts of state accountability systems.

Education is the responsibility of state governments, and states have gone in a variety of directions in the regulation, funding, and operation of their schools. As a result, it is difficult to assess the impacts of individual policies without dealing with the potential impacts of coincidental policy differences.[1]

The basic analysis focuses on growth of student achievement across grades.[2] If the impacts of stable state policies enhance or detract from the educational process in a consistent manner across grades, concentrating on achievement growth implicitly allows for stable state policy influences and permits analysis of the introduction of new state accountability policies.

[1] Hanushek, Rivkin, and Taylor (1996) discuss the relationship between model specification and the use of aggregate state data. The development here builds on the prior estimation in Hanushek and Somers (2001) and the details of the model specification and estimation can be found there.

[2] Here we summarize the results of the analysis in Hanushek and Raymond (2003a, 2003b).

The NAEP testing measured math performance of fourth-graders in 1992 and 1996 and of eighth-graders 4 years after each of these assessments. While the students are not matched, following the same cohort acts to eliminate a variety of potentially confounding achievement influences. We also supplement the raw NAEP data by considering differences in parental education levels and in school spending across these states. Our analysis of achievement relies on growth in achievement in reading and math between fourth- and eighth-graders over the relevant 4-year period, e.g., growth in achievement from fourth grade in 1996 to eighth grade in 2000. Our sample is all states for which the relevant NAEP scores are available.

The potential effects of accountability systems clearly depend on when and where these systems were introduced. Table 1 describes the time path of introduction of accountability systems across states by reference to the length of time that accountability systems have been operating in different states. For these purposes, we define accountability systems as those that relate student test information to schools and either simply report scores or provide rewards and sanctions.[3] By looking at accountability systems in 1996, it is clear that much of the movement to accountability is very recent. In 1996, just 10 states had already introduced active accountability systems, while by 2000 only 13 states had yet to introduce active systems.[4]

We rely on statistical analyses of differences in NAEP growth across states to infer the impact of introducing state accountability. Because a differing set of about 40 states participated in the NAEP testing in each of the years, the amount of evidence is limited. Nonetheless, state accountability systems uniformly have a significant impact on growth in NAEP scores, while other potential influences—spending and parental education levels—do not.

Figure 1 summarizes the impact of existing state systems by tracking the gains in mathematics between 1996 and 2000 for the typical student who progresses from fourth to eighth grade under different systems. These expected gains, calculated from regression analyses of scores on NAEP, illustrate the impact of testing and reporting across states.[5] States were classified according to the type of accountability system they had in place at the time of the NAEP test. (A state's classification could change between the two test years if its accountability system had been newly adopted or changed in the interim.) The typical student in a state without an accountability system of any form would see a 0.7 percent increase in the proficiency scores between fourth and eighth grades. States with "report card" systems display test performance and other factors but do not attach sanctions and rewards to the information. In many ways, these systems simply serve a public disclosure function. Just this reporting moves

[3] We do not include states that place rewards or sanctions ("high-stakes") just on students, for example through use only of a required graduation exam. The school accountability systems are most relevant for No Child Left Behind, but this restriction introduces some differences between our analysis and the analysis of Amrein and Berliner (2002) that is analyzed below.

[4] In all analyses, the universe includes 50 states plus the District of Columbia. Nonetheless, not all states participate in the NAEP exams each year, and the samples fall to around 35 in each year.

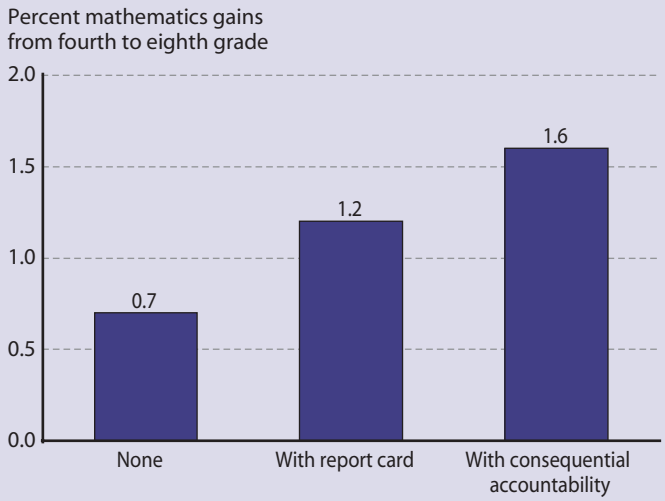
[5] The details of these estimates can be found in Hanushek and Raymond (2003a). The results pool data on NAEP mathematics gains over both the 1992–96 and 1996–2000 periods.

Table 1. Distribution of states with consequential accountability or reporting system: 1996 and 2000

	Number of states	
	1996	2000
No system	41	13
System in place	10	38

NOTE: Distribution includes Washington, DC.
SOURCE: Fletcher and Raymond (2002).

Figure 1. Estimated effects of state accountability systems on gains between fourth grade and eighth grade for National Assessment of Educational Progress (NAEP) mathematics scores: 1996–2000



SOURCE: Author calculations from Hanushek and Raymond (2003a, 2003b).

the expected gain to 1.2 percent. Finally, states that provide explicit scores for schools and that attach sanctions and rewards (what we call "consequential accountability" systems) obtained a 1.6 percent increase in mathematics proficiency scores. In short, testing and accountability as practiced have led to significant gains in student performance over that expected without formal systems.

A complementary analysis by Carnoy and Loeb (2002), while not considering the timing of the introduction of accountability, includes a rating of the stringency of the accountability system that is finer grained than the two categories we employ. It also adds information about student stakes and accountability. Carnoy and Loeb's findings reinforce the present analysis that accountability increases NAEP performance. A variety of other systematic studies of accountability systems within states and local school districts have also investigated what happens when accountability systems are introduced. While we describe the evidence in detail elsewhere (Hanushek and Raymond 2003a, 2003b), it generally supports two conclusions. First, improvements in available measures of student performance occur after the introduction of an accountability system. Second, other short-run changes—such as increases in test exclusions or explicit cheating—are observed. In other words, some unintended consequences often tend to accompany the introduction of accountability, although as of now there is little evidence suggesting that these influences continue over time.

We ourselves have looked explicitly at state differences in special education placement rates and whether they are related to accountability systems. For the period 1995–2000, a time of large change in the use of accountability systems, we see no evidence that increased special education placement is a reaction to accountability systems (Hanushek and Raymond 2003a, 2003b). This analysis does, however, show why some could mistakenly conclude that accountability has an impact: overall special education placement increases within states over this time period, so the introduction of accountability systems in the middle of the period can look like it influences placement.

Carnoy and Loeb (2002) also investigate the impacts of accountability on grade retention and graduation. They demonstrate that there is no discernible negative effect on retention and graduation.

The set of scientific studies of accountability has been presented at a range of scientific conferences, and many have undergone peer review for journal publication. In fact, because of the importance of the topic, the Kennedy School at Harvard held an entire conference on accountability in June 2002, and Brookings published the papers in 2003 (Peterson and West 2003).

The Allure of Counter-Evidence

In late 2002, Amrein and Berliner, hereafter AB, produced a study on the impact of high-stakes accountability systems that garnered considerable attention (AB 2002).[6] Their analysis of 28 states considers the effects on state-specific NAEP scores and college entrance examination measures in the period following adoption of a high-stakes accountability program.[7] Their analysis concludes "there is inadequate evidence to support the proposition that high-stakes tests . . . increase student achievement" (AB 2002, p. 57). The press release that describes the report goes further: "The Berliner-Amrein analyses suggest that, as indicated by student performance on independent measures of achievement, high-stakes tests may inhibit the academic achievement of students, not foster their academic growth."

Because of the importance of the topic, the Kennedy School at Harvard held an entire conference on accountability in June 2002.

A closer look at the research, however, shows it to be fatally flawed both in design and in execution, rendering the conclusions irrelevant. We consider only the effects of accountability systems on NAEP scores in the 26 states that AB record as having adopted grade school high-stakes tests.[8]

It is difficult to ascertain from the main text or the technical appendixes exactly what procedures and definitions AB employed. AB's methodology seems best described as a "pseudo-trend analysis" with, at times, absent baseline data.[9] Given the fact that state-level NAEP data on the math and reading tests are available only for at most four data points, AB essentially were confined to performing case studies of individual states.[10] They purport to examine the change in scores before and after the accountability system was adopted in each state—thus using each state as a control for itself. To give some independent context for these differences, it appears they also generally compared the state change to the change that was observed for the nation as a whole. States were coded as increasing on a particular test if the gains in average test results exceeded the national average change, or coded as decreasing in the opposite case. Finally, all scores were then considered in relation to the relative

[6] This study is described as having been completed for the Great Lakes Center for Education Research and Practice, a Michigan-based think tank. That organization, which is solely financed by National Education Association State Education Affiliate Associations from Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin, in turn describes a key element of its mission as being to "connect with like-minded organizations to partner on key education initiatives."

[7] We have not assessed their identification and timing of high-stakes testing, which apparently can relate both to school stakes and individual student stakes.

[8] Georgia and Minnesota only adopted high school exit requirements, the subject of AB's technical appendix. There are also strong reasons to question their analysis of high school level performance, given the looser degree of correspondence between high school exit requirements and college entrance test results, but that discussion necessarily gets into other issues and only distracts from the key linkages to state accountability that we emphasize here.

[9] For example, they most frequently say in the write-ups for individual states things like "After stakes were attached to tests in Maryland, grade 4 math achievement decreased" (p. 28). But, since fourth-grade NAEP scores in Maryland, like those in *all* of their high-stakes tests except Delaware in 1992–96, increased in every test year, we infer that they really meant to describe a comparison with the average national changes.

[10] Note that reading and math were tested in different years during the 1990s and that many states did not participate in all four waves of NAEP testing.

change in exclusion rates between the national average and the individual states. Where states' exclusion rates exceeded the national average, AB hypothesize that scores should rise because of these exclusions. Thus, whenever exclusion rates moved in the same direction as the observed NAEP test results, they considered the score change contaminated (regardless of the magnitudes involved) and eliminated the state from further consideration as "Unclear."[11] Finally, among states that remained (between 8 and 12 depending on the particular NAEP test), they examined the proportion of states with increases versus those with decreases relative to the national average. Based on this approach, they concluded that "67 percent of the states posted overall decreases in NAEP math grade 4 . . . 63 percent of the states posted increases in NAEP math grade 8 . . . and 50 percent of the states posted increases in NAEP reading grade 4 as compared to the nation after high-stakes tests were implemented." (AB 2002, p. 56)

AB violate the first principle of social science research—the need to control for the condition of interest. They used the 26 states with high-stakes accountability systems and limited their analysis to those states alone. The natural comparison group, however, is the states that had not adopted accountability systems. Such a comparison, which offers some insights into the impact of high-stakes testing as opposed simply to variations among states with high-stakes systems, yields starkly different results than their suggested interpretations.[12] In fact, their results are completely reversed, putting the evidence in line with that previously discussed.

Table 2 simply compares fourth- and eighth-grade NAEP test score gains for the states AB identify as implementing high-stakes testing with those that were not so identified.[13] For either the entire 1992–2000 period or the later period of 1996–2000, the average gain in math for high-stakes states significantly exceeds that for the remaining tested states. The difference in performance is always statistically significant at conventional levels (a nuance that AB never even mention in their 236 pages of analysis).[14]

AB highlight changes in exclusion rates from test taking as a possible influence on state test scores, and differences in exclusions between high-stakes states and others could influence the performance differentials shown. Indeed, many people have suggested that a

[11] In reality, they do not even appear consistent on this, and they violate their own coding scheme more than once. Take, for example, West Virginia, where they state: "Overall NAEP math grade 4 scores increased at the same time the percentage of students exempted from the NAEP increased. Overall, after stakes were attached to tests in West Virginia, grade 4 math achievement **decreased**." [their emphasis]

[12] While we reproduce their analysis with a larger set of observations, this should not be construed as an endorsement of the analytical approach. More rigorous tools yield more reliable results. We follow their lead in order to show how their answers would have differed had they applied their own approach correctly.

[13] Note that for each of the comparisons data are available for 34 to 36 states with between 18 and 20 being in the AB high-stakes sample. The limited number reflects the varying participation of states in the NAEP testing.

[14] Statistical testing is done to guard against changes in test performance that simply reflect random score differences that do not represent true differences in student performance. Such random differences could, for example, reflect chance differences in the tested population, small changes in question wording, or events specific to the testing in a given year and given state. In their subsequent defense of their analysis, AB assert that such testing is unnecessary and may even be inappropriate, but this assertion is obviously incorrect (AB 2003).

Table 2. Average gains in National Assessment of Educational Progress (NAEP) mathematics scores, by Amrein-Berliner (AB) high-stakes states versus other states: 1992–2000

	Change in fourth-grade NAEP mathematics scores		Change in eighth-grade NAEP mathematics scores	
	1992–2000	1996–2000	1992–2000	1996–2000
AB high-stakes states	9.2	4.2	8.8	4.5
Other states	3.8	2.3	4.0	1.7
High-stakes advantage	5.3	1.9	4.8	2.8
Statistical significance	$p<.001$	$p<.04$	$p<.003$	$p<.02$

SOURCE: Author calculations.

consequence of the introduction of high-stakes testing is an increase in test exclusions.

The hypothesized effect of accountability on test exclusions does not appear important in explaining the aggregate accountability results. For the nation as a whole, exclusion rates on the eighth-grade NAEP math tests were the same in 2000 as in 1992, while the fourth-grade exclusions over that time period fell slightly. Table 3 shows evidence for the NAEP exclusion rates for the 1992–2000 period for the high-stakes and non-high-stakes states. While the change in exclusion rates over the 1990s is slightly higher for high-stakes states in the testing of eighth-grade mathematics, it is slightly lower for fourth-grade mathematics when compared to other states. But neither difference in average exclusion by accountability status is statistically significant.

We also standardize the achievement gains for observed changes in exclusions through regression analysis. Interestingly, while changes in exclusion rates are significantly related to changes in eighth-grade scores, they are not significantly related to changes in fourth-grade scores—underscoring the need to analyze central maintained hypotheses. Table 4 compares such adjusted estimates of the achievement gain advantage of high-stakes tests to the previously unadjusted differences. Again, there are small effects on the estimated impact of high-stakes testing on gains, but in all cases states that introduce high-stakes testing outperform those that do not by a statistically significant margin. In sum, the previous estimates are not driven by test exclusions.

AB's choice of the pseudo-trend design is even more mysterious when one considers that it could not be applied squarely to their sample. In eight states—Colorado, Indiana, Louisiana, New Jersey, New Mexico, Oklahoma, Tennessee, and West Virginia—high-stakes testing was identified by AB as having been adopted prior to 1990 or in 2000. Because these adoptions fall outside of the relevant testing period, any pre/post comparison based on NAEP data is impossible. Thus, we refer to their design as "pseudo-trend" because they frequently lack data before or

Table 3. Changes in NAEP mathematics exclusion rates, by Amrein-Berliner (AB) high-stakes states versus other states: 1992–2000

	Change in fourth-grade mathematics exclusion rates		Change in eighth-grade mathematics exclusion rates	
	1992–2000	1996–2000	1992–2000	1996–2000
AB high-stakes states	3.8	1.3	3.4	2.3
Other states	4.1	2.0	2.6	1.9
High-stakes differential	–0.3	–0.7	0.8	0.4
Statistical significance	$p<.76$	$p<.44$	$p<.40$	$p<.64$

SOURCE: Author calculations.

Table 4. Adjusted average gains in NAEP mathematics scores, by Amrein-Berliner (AB) high-stakes states versus other states: 1992–2000

	Change in fourth-grade NAEP mathematics scores		Change in eighth-grade NAEP mathematics scores	
High-stakes advantage	1992–2000	1996–2000	1992–2000	1996–2000
Unadjusted for test exclusions	5.3	1.9	4.8	2.8
Statistical significance	$p<.001$	$p<.04$	$p<.003$	$p<.02$
Adjusted for change in test exclusions	5.2	2.3	3.7	2.5
Statistical significance	$p<.001$	$p<.02$	$p<.02$	$p<.02$

NOTE: Adjusted average gains come from regression of NAEP score changes on exclusion rate changes.
SOURCE: Author calculations.

after the treatment of interest, and they often have just two or three test scores that are not even aligned with the treatment. For some states, they observe only a single test score change, obviously making any pre/post comparison unreliable.

The use of national average changes in NAEP scores as a reference point further confounds the study. Any effect of accountability systems is already captured in the national score change. By 1996, only 10 states had an accountability system in place, so the effect might not excessively affect the average. But by 2000, a majority of states were on board, so their impacts affected the national average change to a much greater degree. Late-adopting states are effectively being compared to other high-stakes states, making it difficult to show relative gains and completely rendering moot the interpretation that any differences reflect the high-stakes treatment. To take a purely hypothetical example, assume that 6 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. AB's approach would say that accountability had failed: just 6 states beat the national average, while 20 were below the average. In fact, ignoring any complications of exclusions, AB 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, of course, would be to compare gains of high-stakes states to those of no-accountability states.

A subtler but important issue arises when the timing of adoption of an accountability system was bracketed by NAEP tests. It is clear that AB did not use a consistent convention. In some cases, it appears that they used the NAEP results from the period immediately prior and immediately following adoption of accountability, but in others, it appears that they used a different time interval, in some cases starting after the accountability systems were adopted. The one consistent choice appears to be reliance on the least flattering results (for high-stakes accountability).

The implications of these nonscientific procedures is best seen within the context of their finding of "harm." Table 5 examines the set of states where AB concluded that fourth-grade NAEP math scores decreased with the introduction of high-stakes testing. For the eight such identified states, we present aggregate information on testing and results. In three of the eight states (New Mexico, Oklahoma, and West Virginia), AB identify the introduction of high-stakes testing as falling outside the testing period (which did not begin until the 1990s). Moreover, no real trend data in math gains are available for Nevada and Oklahoma, where only a single period of test change is observed. During the 1992–96 period when Kentucky, Maryland, and Missouri intro-

Table 5. Data on NAEP fourth-grade mathematics performance in states identified by Amrein-Berliner (AB) as decreasing after the introduction of high-stakes tests: 1992–2000

States where AB declared decreases in NAEP scores	Introduction of high-stakes testing (AB date)	1992–1996	1996–2000	1992–2000
Kentucky	1994	**4.9[2]**	1.0	**5.9[3]**
Maryland	1993	**3.4[2]**	1.6	**5.0[3]**
Missouri	1993	**2.5[3]**	**3.8[2]**	**6.3[3]**
Nevada	1998	**N/A**	**2.7[3]**	**N/A**
New Mexico	**1989[1]**	0.5	0.0	0.6
New York	1999	**4.2[2]**	**3.9[2]**	**8.1[2]**
Oklahoma	**1989[1]**	**N/A**	**N/A**	**4.7[3]**
West Virginia	**1989[1]**	**8.1[2]**	1.5	**9.6[2]**

N/A—NAEP data unavailable for this time period.

[1] No NAEP tests at or before introduction of high-stakes testing.

[2] Change in NAEP scores exceeds the average change in NAEP both for all states and for states not adopting high-stakes testing.

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

NOTE: Bold entries highlight evidence concerns discussed in text.

SOURCE: Author calculations.

duced high-stakes testing, two had math gains exceeding the average for all tested states, and one had gains that just exceeded the average for states that did not introduce high-stakes testing.[15] Nevada, which they record as introducing high-stakes testing in 1998, had gains during 1996–2000 that exceeded gains for non-high-stakes states. Over the entire period of 1992–2000, five of the six states for which data are available showed gains that at least exceeded the average for non-high-stakes tests; New York and West Virginia exceeded the average for all states. And this is the group of states that AB identify as being harmed by high-stakes testing! *Not a single state* provides evidence of harm following the introduction of high-stakes testing. When read correctly, if anything, the evidence points to generally higher performance in this group of states.

The final blow to the credibility of AB's results comes at the point of drawing inferences based on their analysis. Regardless of the choice of design, and ignoring the selective use of NAEP scores, we would still expect AB to consider all the available data *as they had constructed it* to draw conclusions. But they did not. First, they eliminated all information about the magnitude of score changes, relying solely on whether scores increased or decreased. Second, they eliminated all the states that they judged to be "unclear," which reduced the final tally to "improved vs. declined" instead of "improved vs. all states that adopted high-stakes."[16] For instance, they recorded positive or negative results on the NAEP fourth-grade math test for just 12 of the 26 states with high-stakes for grades K–8. AB found that fourth-grade math scores increased at a slower rate than the national average in eight of the remaining states (those in table 5), faster in just four. Yet they write this up in a highly misleading fashion, claiming "67 percent of the states posted overall decreases in NAEP math grade 4 performance as compared to the nation after high-stakes tests were implemented." Actually, AB witnessed gains slower than the national average in just 8 of 26 high-stakes states, or 31 percent.

Instead of concluding that the evidence does not support the proposition that high-stakes accountability increases student achievement, it would be more accurate to say that the chosen evidence by AB does not support any inference at all.

Simply applying the underlying approach of AB to all of the data on NAEP achievement completely reverses their conclusions. High-stakes test states on average perform significantly better than non-high-stakes states. For the reasons described previously, we still do not think that these simple comparisons are the best way to analyze this question, but this analysis demonstrates that there is no difference in the broad results from their crude approaches and the preferred analytical approaches we described previously.

The competing evidence on accountability program performance raises a number of disturbing issues.

Not in a Vacuum

The competing evidence on accountability program performance raises a number of disturbing issues. One is how unaware or indifferent the media and many policymakers are to quality differences in the available evidence. The recent publicity surrounding the AB essay highlights the vulnerability of key public policy initiatives to faulty evidence and badly informed reporting.[17] Distinct from other policy fields, reports in education seem to be taken at face value or—worse—on the political orientations of the authors, independent of the rigor of the analysis or the suitability of the inferences that are drawn. While the most obvious example recently concerned the me-

[15] In terms of what periods were looked at by AB, it is difficult to come up with the rule for decisions on NAEP scores that includes both Maryland and Missouri as "decreasing" states.

[16] As described above, the label "unclear" rests on their strong and untested hypothesis about the impact of exclusion rates on scores. Results are unclear whenever the movement in exclusion rates is the same direction as the movement in test scores, regardless of the magnitude of either change.

[17] Most notable among the publicity was a front page article in the New York Times (Winter [2002]). A link to this article currently appears on the home page for the Great Lakes Center for Education Research and Practice: http://www.nytimes.com/2002/12/28/education/28EXAM.html. Other newspapers and professional publications dutifully provided their own reporting of the AB results.

dia, the problem applies as well to many other actors on the education landscape, including the legislative and executive leadership in many states.

The issue of evidence quality is of prime importance when individuals serve a gatekeeper function for disseminating information to the general public. The media acts as a filter to select issues that merit attention and then distills them into a few key points. Decisionmakers in education agencies serve a similar function when they attempt to reflect the effectiveness of the programs they have implemented. Individuals in these positions are trusted, and expected, to go beyond the press release or a superficial examination of a report or analysis by checking the facts, gauging the credibility of the analytic approach, and vetting the results. We would certainly expect this if the topic under investigation were an allegation of fraud or a new breakthrough in power generation. We need similar assurances in education.

Perhaps this disregard is understandable when one considers that the issue of the quality of evidence has only recently been raised among educators themselves. A recent National Research Council panel was convened to assess "scientific principles for education research"—a type of inquiry unheard of in other research and policy fields (Shavelson and Towne 2002). Most schools of education offer courses in research methods as part of the curriculum, but a wide variety of techniques are taught, determined in no small part by the training and interests of the faculty teaching the courses and not limited to traditional scientific inquiry. This is not to say that there are no appropriate uses for the variety of analytic skills that are taught. However, when significant public policies involving many millions of dollars are on the line, as in the case of school and student accountability programs, evidence must meet the highest scientific standards. The analysis should be rigorous enough to consider and objectively to control for potentially competing explanatory factors, and the resulting evidence must be reliable. It should not be argued on political grounds masquerading as science.[18]

When significant public policies involving many millions of dollars are on the line, evidence must meet the highest scientific standards.

Federal policy has also taken a turn towards more stringent standards of evidence. The No Child Left Behind Act includes strong requirements for employing educational programs based on solid scientific research. The creation of the Institute of Education Sciences is clearly directed at improving the quality of educational research. And, reacting to obvious quality concerns about research that was being used to support policy, the U.S. Department of Education in 2002 funded the What Works Clearinghouse to establish strict scientific criteria for studies on program performance. In an effort to provide a "trusted source of scientific evidence," the Clearinghouse is designed to concentrate primarily on the quality of the research design and the rigor of the analytic techniques. (See http://w-w-c.org)

Reporters should not be expected to be experts in statistical analysis any more than they are expected to be fully versed in biochemistry or investment banking regulations. But it is not unreasonable to hold up a standard of reasonable scrutiny (bringing in expertise if needed as is done for medical and scientific reporting).

It is also not as if the issue is unimportant. Improving our educational performance would arguably lead to greater gains for society than any of the medical breakthroughs of the past decade. For example, had there been true educational improvements following *A Nation at Risk*—putting U.S. student achievement on par, say, with that of students in better performing European countries—it has been estimated that the GDP of the United States would have expanded sufficiently by 2002 to pay for all K–12 expenditures.[19]

[18] See the debates about the effectiveness of accountability systems that entered into the 2000 presidential elections; Grissmer et al. (2000), Klein et al. (2000), and Hanushek (2001).

[19] See Hanushek (2003a, 2003b) (http://www.educationnext.org/20032/index.html).

What We Do Not Know

We have suggestive evidence that accountability as implemented in the 1990s has been helpful. It is clear that, for one reason or another, performance has been better in accountability states than in nonaccountability states. We also have evidence that a number of unintended consequences have followed the introduction of accountability. We do not wish to suggest that we yet have anywhere near the amount of reliable evidence that is needed for developing fully satisfactory testing and accountability systems. But this is far different from completely retreating from assessing and reporting schooling outcomes.

The findings leave us short of what we would like to know for policy purposes.[20] We do not understand how best to design accountability systems that can be directly linked to incentive systems. For example, the vast majority of state accountability systems report average performance for each school on various state tests. These are sometimes disaggregated for, say, race and ethnic groups. But, because these average scores are highly dependent on factors outside the control of schools—such as families and friends—it would not be appropriate to base school performance rewards on these unadjusted average scores. Doing that would encourage schools to concentrate more on who is taking the test than on how their scores can be improved. Incentives are best attached to the value-added for which schools and teachers are responsible.

Similarly, uncertainty remains about the best set of tests to measure accomplishment of the learning standards of each state. Concerns about any possible narrowing of the curriculum or inappropriate changes in instructional practice are in large part concerns about the quality of the testing—because the entire intent of the accountability systems is that teachers do in fact teach to a well-designed set of tests that adequately reflect the range of material that students should know.

Federal legislation in the No Child Left Behind Act represents an important starting point in a process to improve the performance of our schools. It established the necessity for regular annual testing of students and the public reporting of results. It also made some guesses about how to build incentives and requirements into the system. The hope (and intent) of the anti-accountability forces is that regular testing and reporting be nipped in the bud. The challenge to everybody is ensuring that we learn about accountability and adjust any current flaws before the anti-accountability forces succeed. Their success would surely leave our children and our nation worse off.

[20] Issues of accountability system design and of incentive aspects of accountability systems are discussed in Hanushek and Raymond (2003b). These analyses also assess the available evidence on various design issues.

References

Amrein, A.L., and Berliner, D.C. (2002). *The Impact of High-Stakes Tests on Student Academic Performance: An Analysis of NAEP Results in States With High-Stakes Tests and ACT, SAT, and AP Test Results in States With High School Graduation Exams*. Tempe, AZ: Educational Policy Research Unit, College of Education, Arizona State University.

Amrein, A.L., and Berliner, D.C. (2003). Does Accountability Work? [Correspondence]. *Education Next*, *3*(4): 8.

Carnoy, M., and Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis, 24*(4): 305–332.

Fletcher, S.H., and Raymond, M.E. (2002, April). *The Future of California's Academic Performance Index.* Report to the California Secretary of Education. Stanford, CA: CREDO, Hoover Institution, Stanford University.

Grissmer, D.W., Flanagan, A., Kawata, J., and Williamson, S. (2000). *Improving Student Achievement: What NAEP State Test Scores Tell Us*. Santa Monica, CA: RAND.

Hanushek, E.A. (2001). Deconstructing RAND. *Education Matters, 1*(1): 65–70.

Hanushek, E.A. (2003a). The Importance of School Quality. In P.E. Peterson (Ed.), *Our Schools and Our Future: Are We Still At Risk?* (pp. 141–173). Stanford, CA: Hoover Institution Press.

Hanushek, E.A. (2003b). Lost Opportunity. *Education Next*, *3*(2): 84–87.

Hanushek, E.A., and Raymond, M.E. (2003a). Improving Educational Quality: How Best to Evaluate Our Schools? In Y. Kodrzycki (Ed.), *Education in the 21st Century: Meeting the Challenges of a Changing World* (pp. 193–224). Boston: Federal Reserve Bank of Boston.

Hanushek, E.A., and Raymond, M.E. (2003b). Lessons About the Design of State Accountability Systems. In P.E. Peterson and M.R. West (Eds.), *No Child Left Behind? The Politics and Practice of Accountability* (127–151). Washington, DC: Brookings.

Hanushek, E.A., Rivkin, S.G., and Taylor, L.L. (1996). Aggregation and the Estimated Effects of School Resources. *Review of Economics and Statistics, 78*(4): 611–627.

Hanushek, E.A., and Somers, J.A. (2001). Schooling, Inequality, and the Impact of Government. In F. Welch (Ed.), *The Causes and Consequences of Increasing Inequality* (pp. 169–199). Chicago: University of Chicago Press.

Klein, S.P., Hamilton, L.S., McCaffery, D.F., and Stecher, B.M. (2000). *What Do Test Scores in Texas Tell Us.* Santa Monica, CA: RAND.

Peterson, P.E., and West, M.R. (Eds.). (2003). *No Child Left Behind? The Politics and Practice of Accountability*. Washington, DC: Brookings.

Shavelson, R.J., and Towne, L. (Eds.) (2002). *Scientific Research in Education*. Washington, DC: National Academy Press.

Winter, G. (2002, December 28). Make-or-Break Exams Grow, but Big Study Doubts Value. *New York Times*, p. A1.