Grinding the Antitesting Ax

More bias than evidence behind NRC panel's conclusions

Incentives and Test-Based Accountability in Education

A report from the
National Research Council

Checked by Eric A. Hanushek

PHOTO / SHUTTERSTOCK.COM/ DIGITALREFLECTIONS

The No Child Left Behind Act of 2001 (NCLB) was scheduled for reauthorization in 2007, and its future has in recent months garnered renewed attention. Yet so far, Congress has found it impossible to reach sufficient consensus to update the legislation, as competing groups want to a) keep all the essential features of the current law as a way of maintaining the pressure on schools to teach all students, b) modify the federal law by moving to a value-added or some alternative testing and accountability system, or c) eliminate federal testing and accountability requirements altogether, reverting to the days when the compensatory education law was simply a framework for distributing federal funds to school districts. Critics of NCLB's testing and accountability requirements have a litany of complaints: The tests are inaccurate, schools and teachers should not be responsible for the test performance of unprepared or unmotivated students, the measure of school inadequacy used under NCLB is misleading, the tests narrow the curriculum to what is being tested, and burdens imposed upon teachers and administrators are excessively onerous.

But in all the acrimonious discussion surrounding NCLB, surprisingly little attention has been given to the actual impact of that legislation and other accountability systems on student performance. Now a reputable body, a committee set up by the National Research Council (NRC), the research arm of the National Academy of Sciences, has reached a conclusion on this matter. In its report, *Incentives and Test-Based Accountability in Education*, the committee says that NCLB and state accountability systems have been so ineffective at lifting student achievement that accountability as we know it should probably be dropped by federal and state governments alike. Further, the committee objects to state laws that require students to pass an examination for a high school diploma. There is no evidence that such tests boost student achievement, the committee says, and some students, about 2 percent, are not getting their diplomas because they can't—or think they can't—pass the test. The headline of the May 2011 NRC press release is frank and bold in the way committee reports seldom are: "Current test-based incentive programs have not consistently raised student achievement in U.S.; Improved approaches should be developed and evaluated."

Needless to say, the report can be expected to play an important role in the continuing debate over NCLB. Upon its initial release, the report captured top billing, appearing on *Education Week*'s front page. Certainly, the NRC intends for the report to influence the NCLB conversation, rushing a draft version to the media five months before the completed report was available to the public.

Unfortunately, the NRC's strongly worded conclusions are only weakly

supported by scientific evidence, despite the fact that NRC's stated mission is "to improve government decision making and public policy, increase public understanding, and promote the acquisition and dissemination of knowledge."

The Report

Reports from the NRC are generally treated as highly credible. The NRC convenes panels of outside experts who volunteer their time to provide consensus opinions on issues of policy significance. And this particular panel includes a number of especially qualified researchers (see sidebar). The committee chair, Michael Hout, is a member of the National Academy of Sciences; 7 of the 17 panel members have named professorships; 2 are deans (of law and education schools); and a majority have published articles about testing, accountability, or incentives.

When it comes to gathering together the general literature, both theoretical and empirical, on the use of incentives in various contexts, the committee's work is solidly constructed. But this strong scientific discussion of theory and empirical analysis of incentives and accountability breaks down when it comes to the committee's core purpose: evaluating accountability regimes in education that employ incentives and tests.

The report comes to two policy conclusions: NCLB and state accountability systems have proven ineffective and state-required high-school exams are counterproductive. The unequivocal presentation of the conclusions is clearly designed to leave little doubt in the minds of policymakers. When the underlying evidence is examined, however, it becomes apparent that neither conclusion is warranted. Instead of weighing the full evidence before it in the neutral manner expected of an NRC committee, the panel selectively uses available evidence and then twists it into bizarre, one might say biased, conclusions.

Selecting Evidence

To get a grasp of the bias that motivated the report's authors, consider how its first conclusion is phrased:

> Test-based incentive programs, as designed and implemented in the programs that have been carefully studied, have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries.

Note especially that the conclusion does not say that there is no evidence that testing and accountability work. It says that testing and accountability, by themselves, cannot lift the United States to the level of accomplishment reached by the world's highest-achieving countries, an extraordinary standard for evaluating a policy innovation. To catch up to the leading countries would require gains of at least half of a standard deviation, or roughly two years of learning (see "Are U.S. Students Ready to Compete?" *features*, Fall 2011). No individual reform on the public agenda—neither merit pay, class size reduction, salary jumps for teachers, nor Race to the Top—can claim or even hope for anything close to that level of impact. The appropriate question is not whether testing and accountability is a panacea, but whether it has proven worthwhile.

Members of the Committee on Incentives and Test-Based Accountability in Public Education, Division of Behavioral and Social Sciences and Education, National Research Council

Michael Hout (Chair), Department of Sociology, University of California, Berkeley

Dan Ariely, Fuqua School of Business, Center for Cognitive Neuroscience, and School of Medicine, Duke University

George P. Baker III, Harvard Business School

Henry Braun, Lynch School of Education, Boston College

***Anthony S. Bryk** (member until 2008), Carnegie Foundation for the Advancement of Teaching

Edward L. Deci, Department of Psychology, University of Rochester

Christopher F. Edley, Jr., School of Law, University of California, Berkeley

Geno Flores, California Department of Education

Carolyn J. Heinrich, LaFollette School of Public Affairs, University of Wisconsin-Madison

Paul Hill, Center on Reinventing Public Education, University of Washington

***Thomas J. Kane** (member until February 2009), Graduate School of Education, Harvard University, and Bill & Melinda Gates Foundation, Seattle, Washington

Daniel M. Koretz, Graduate School of Education, Harvard University

Kevin Lang, Department of Economics, Boston University

Susanna Loeb, Graduate School of Education, Stanford University

Michael Lovaglia, Department of Sociology, University of Iowa, Iowa City

Lorrie A. Shepard, School of Education, University of Colorado, Boulder

Brian Stecher, RAND Corporation, Santa Monica, California

Staff

Stuart W. Elliott, Study Director

* Bryk and Kane did not participate in the final deliberations.

SOURCE: National Research Council, *Incentives and Test-Based Accountability in Education*, available on the National Academies Press web site (www.nap.edu)

By that more appropriate standard of judgment, the committee's own data indicate that testing and accountability have proven effective, if not quite the spectacular success promised by those who enacted NCLB into law. The committee report tells us that the average estimated impact of these interventions is 0.08 standard deviations of student achievement. In other words, the average student in a state without accountability would have performed at the 53rd percentile of achievement had that student been in a state with an accountability system, all other things being equal.

That estimate may well be too low. The report states that "our literature review is limited to studies that allow us to draw causal conclusions about the overall effects of incentive policies and programs," and then it goes on to describe several types of studies that would be excluded by this criterion. Where does the 0.08 come from? The committee considers a review from 2008 of 14 studies, and 4 studies conducted after that review. The review presents an average impact of 0.08. The NRC committee apparently felt no need to look any further and ignored the fact that a majority of the 14 studies would not come close to meeting its standard of enabling a "causal conclusion." The committee determines that one of the more recent studies also supports an estimate of 0.08, although that study's authors prefer estimates that are much higher. The 14 earlier studies and the 4 later ones produce a wide distribution of estimated impacts, but the committee makes no attempt to investigate whether the unusual estimates suggest circumstances under which accountability seems particularly effective (or ineffective). The committee chooses to emphasize the studies with negative findings (10 percent) while downplaying a number of those that have positive findings (90 percent). Thus the NRC mantra, repeated with slightly different wording throughout the report: "Despite using them for several decades, policymakers and educators do not yet know how to use test-based incentives to consistently generate positive effects on achievement and to improve education." Apparently, the inconsistent results heralded in the press release reflect the 10 percent of studies that differed from the overwhelming majority.

The committee chooses to emphasize the [accountability] studies with negative findings (10 percent) while downplaying a number of those that have positive findings (90 percent).

Small Gains Add Up

Let us put this concern aside and consider the increment in student performance of 0.08 standard deviations of individual achievement that the committee presents as its best estimate. Is that so small an effect that it cannot justify continuation of testing and accountability? Consider that this is the average effect of a program that has been implemented on a national scale, affecting students across the country. We are hard pressed to come up with *any* other education program working at scale that has produced such results. Moreover, these average gains are the result of accountability systems that many people believe have important flaws. Even larger gains might be expected if those flaws could be corrected, as many experts, though not the NRC panel, have suggested.

The estimated benefits from a 0.08 standard deviation gain in student performance vastly outweigh its estimated costs. The cost of designing, administering, grading, and reporting the results from statewide examinations have been estimated at between $20 and $50 per pupil, a trivial sum considering that per-pupil education expenditures in the United States run above $12,000 annually. Most reforms—including class size reduction, merit pay, across-the-board raises for teachers, in-service training programs, or the scaling up of charter schools—would cost many, many times as much. For these innovations to have the same kick for every dollar invested, results would have to be improbably large.

The NRC, instead of considering these actual costs, suggests that implicit costs in the form of narrowed curricula are the most important, but it provides no evidence for its view.

What might the economic impact of a 0.08 standard deviation improvement in average achievement nationwide be? Along with University of Munich professor Ludger Woessmann, I have estimated the impact on U.S. Gross Domestic Product (GDP) of higher levels of student achievement. These estimates project the historical pattern of growth to determine the result of gains in student achievement, calculate the additions to GDP over the next 80 years, and discount them back to today so that they are comparable to other current investments. A 0.08 improvement has a present value of some $14 trillion, very close to the current $15 trillion level of our entire GDP, and equivalent to $45,000 for every man, woman, and child in the U.S. today. In other words, an inexpensive program that affects every student nationwide can, over the long run, have a very large impact, even if its average effect seems at first glance to be quite small. Indeed,

if we figured testing cost $100 per student each year for the next 80 years and we tested all students rather than the limited grades tested now, the rate of return on the investment would be 9,189 percent. Google investors would be envious.

Several omissions from the report are also noteworthy. The report gives only passing attention to the positive impact of NCLB on the education of the most disadvantaged students, a consequence of the requirement to report performance by specific subgroups (e.g., racial and ethnic groups and the economically disadvantaged). The NRC report's main reference to this feature of current accountability systems is that consideration of subgroup performance has added analytical difficulties because of the smaller samples.

Perhaps more telling, this panel of experts on testing and incentives makes absolutely no effort to describe how accountability programs could be improved. Being good researchers themselves, they do favor continued research on testing, however, and provide recommendations on what research should be done, which not surprisingly matches their own interests and expertise.

Lower the Bar?

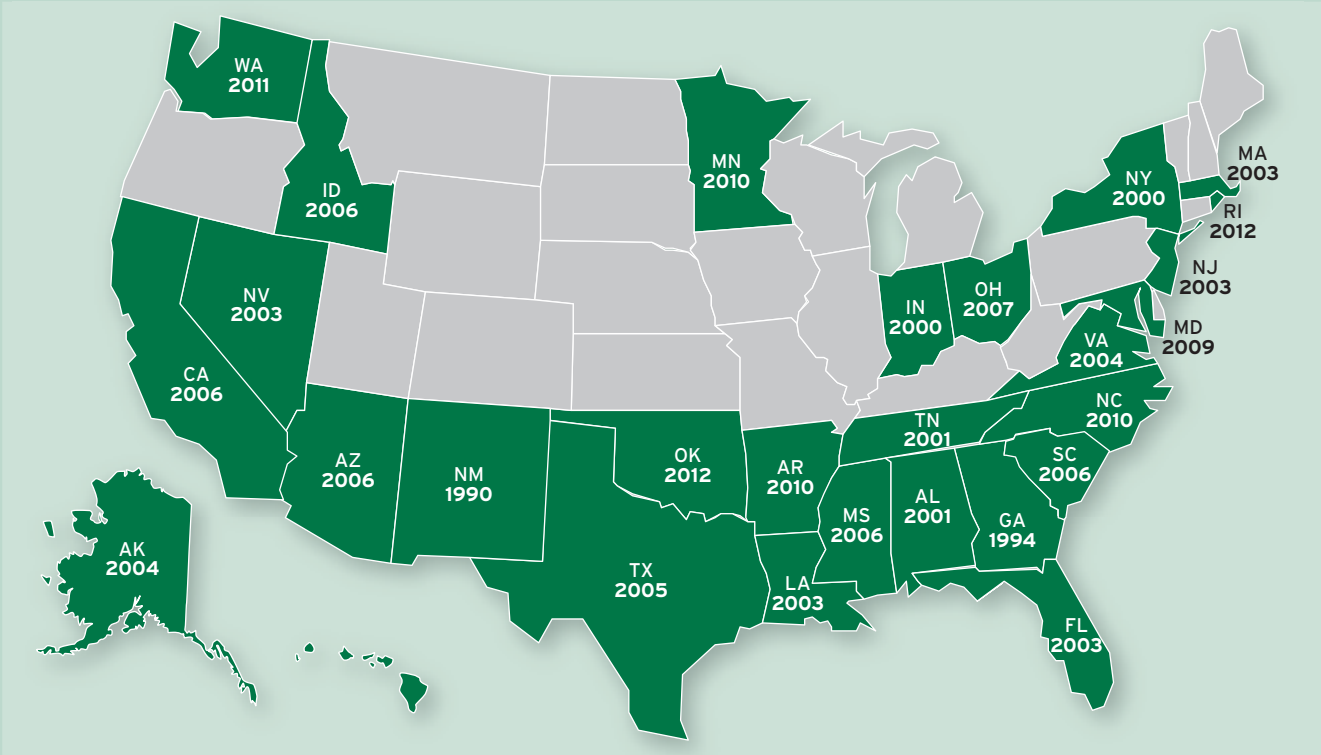
The report also addresses a second, widely used accountability policy: high-school exit exams that hold students responsible for meeting a set of content standards. The report's second conclusion reads,

> The evidence we have reviewed suggests that high school exit exam programs, as currently implemented in the United States, decrease the rate of high school graduation without increasing achievement.

The panel strongly suggests that states that impose an exit exam should repeal this requirement. To understand this conclusion, it is necessary to understand the exams themselves and to evaluate the evidence behind the committee's conclusion.

State Expectations (Figure 1)

Currently 28 states include passing an exit exam among criteria for receiving a high school diploma, with almost all having instituted the policy in the last decade.



Note: Date indicates when current requirement came into effect.

SOURCE: Center on Education Policy, "State High School Tests: Exit Exams and Other Assessments," December 2010 (www.cep-dc.org)

Currently, more than half of the states require that students pass a test of some sort to obtain a normal diploma (see Figure 1), and virtually all of these current requirements have been put in place since 2000. The tests almost always cover English and math, but many states add science and history. Test difficulty varies by state, but the modal level is grade 10. Although that standard may seem low, it is considerably more stringent than the standards that existed prior to 1990, when no state had a test reaching even the 9th-grade level. The current tests are not as high a barrier to high school graduation as they are often alleged to be, as a student may generally take the exam multiple times in order to achieve a passing score. And in all but three states (South Carolina, Tennessee, and Texas), students can either appeal the test result, if they feel the score misrepresents their accomplishments, or obtain a diploma by some alternative path.

The motivations for administering exit exams are to create incentives for students to apply themselves to the task of learning and to set uniform (minimum) quality standards for the state's schools. Such content standards provide guidelines to schools about what to teach. They also indicate to colleges and universities what knowledge and abilities a graduate can be expected to possess. And they give similar information to prospective employers.

According to the best available evidence (discussed below), perhaps 2 percent of students are induced to drop out of school either because of failure to pass the exam or because of fear of not being able to pass the exam. Implicitly, the committee assumes this consequence does considerable harm to the affected students, given the substantial economic rewards that accrue, on average, from receiving a high school diploma. But average effects do not necessarily apply to the 2 percent on the border line between graduating and failing to graduate from high school. The impact for this particular group of students is likely to be much less, unless you make the bizarre assumption that it is only the diploma—not what the student learns—that affects job prospects and future income. The people who are induced to drop out because they cannot pass a 10th-grade exam would most likely be near the bottom of the earnings distribution of graduates were they to be handed a diploma. The economic impact on these students will be much lower than the average difference between graduate and dropout.

Perhaps the best argument against exit exams is simple: If a student shows up for school for 12-plus years and cannot pass a 10th-grade exam, it must be the school's fault, and it would be unfair to hold the student responsible. This argument, interestingly enough, is the precise opposite of one of the primary arguments against the testing and accountability provisions of NCLB: We should not hold schools responsible for low achievement, because achievement is affected by student motivation and family background characteristics beyond the school's control. Taken together, the arguments embedded in the committee's two conclusions imply that nobody—not schools, not teachers, not even students themselves—bears responsibility for low student achievement.

Interestingly, the committee's conclusion with respect to exit exams does not pick up on the full report's emphasis on the importance of the design features of incentive systems, which include warnings that tests aimed at ensuring minimum competency may lower expectations, and concerns about both the potential narrowing of the curriculum and the tendency for score inflation on a known test. Instead, the presumed problem is the inherent unfairness of denying a diploma to a student who has met the attendance and course distribution requirements for a diploma.

If the main objective is to maximize high school graduation, there are many ways to do that. We could eliminate all exams, even those administered by teachers. We could loosen up course requirements. We could offer the diploma after 10 or 11 years of schooling, instead of 12. Of course, nobody is willing to take such steps, even though class exams, course requirements, and the inclusion of the 12th grade of schooling all have negative impacts on graduation rates. So why then does the NRC promote the idea of eliminating a 10th-grade-level examination as a requirement for high school graduation on the narrow basis that a few students will, as a result, not earn the degree? Is the NRC also against the movement of many states toward increasing the required amount of math or moving to college and career-ready standards?

The Data Shuffle

Let's examine the evidence the committee supplies for its exit exam conclusion. The report marshals three studies that explore the issue: two on dropouts and one on achievement. Evaluating the impact of exit exams on achievement is inherently difficult. Because the exams apply to everybody in a state at the same time, it is not possible to compare students of the same age within the same state to find out the impact of exams. It is possible, however, to look at different cohorts of students, for example, those who attended school before the exam was in place and those who attended after, and to compare these to similar cohorts in other states where no such change in policy took place. In conducting this type of study, one must rule out other differences, such as those in family background or those in state education policies that might also affect student performance over time. Even when these challenges are met, one cannot be entirely sure of the results, as exit exams may influence student and school performance even before they come into effect, if teachers and

students know that they will soon be introduced, which is usually the case.

The committee tosses out every exit-exam study (save three) that has ever been conducted on the grounds that it is not possible "to draw causal conclusions about the overall effects of test-based incentives" (that is, the very same criteria the committee ignored in considering school-level accountability). Some of the excluded studies use the well-regarded quasi-experimental technique known as regression discontinuity analysis. In the committee's view, "Such regression discontinuity studies provide interesting causal information about the effect of being above or below the threshold, but they do not provide information about the overall effect of implementing an incentives program." That criticism is odd, since the impact of an exit examination is of special interest for exactly those students on the cusp of adequate levels of achievement. While these excluded studies are not really appropriate for studying achievement, they tend to show little impact of exit exams on dropout behavior or graduation outcomes.

The committee relies for its conclusion regarding exit examinations exclusively on a 2009 study by Eric Grodsky, John Robert Warren, and Demetra Kalogrides. Because of the significance of this piece of research for the committee project as a whole, it is worth considering in some depth. The Grodsky team identified trends in student achievement in each state that administers an exit examination by drawing on data provided by the long-term trend assessments of the National Assessment of Educational Progress (NAEP). The long-term NAEP, begun in the late-1960s and continued with testing every few years, was designed to provide consistent score information to judge achievement of the nation as a whole. It was not designed to be used to evaluate the schools of any particular state or district. As a result, NAEP never collected in its long-term trend assessment a representative sample of students for any specific state, and the median number of tested students in each state was very small.

Grodsky et al. pretend that the NAEP provides them with just that: a representative sample of students for each state. They assume that the average performance of students in each state on the long-term NAEP provides an accurate measure of the average performance of students in that state, thereby violating the first principle of statistical sampling.

The message that comes through is clear: keep working on test development, but never use tests for any incentive or policy purposes.

They then merge the information with information on the timing of the adoption of an exit exam by a state between 1971 and 2004. The study includes observations of math and reading achievement at 9 and 10 different points in time, respectively. The researchers report results for achievement of 13-year-olds and 17-year-olds separately, acknowledging that there are limitations to using either cohort. Thirteen-year-olds may be too young to detect the impact of exit exams, while the sample of 17-year-olds suffers from the noninclusion of school dropouts.

The Grodsky analysis encounters a further difficulty. For the most part, the researchers consider only the very early years, when exit exams were first introduced, a time when the exams were set at a very low level of difficulty, below that of a 9th-grade student. Only 1 percent of the observations included in their analysis are for states that had an exit exam rated at the 9th-grade level or higher, as most current examinations are.

Not only does the Grodsky team rely on inadequate data, but the analysis itself is flawed. Any attempt to see the effects of state tests should compare the changes that occur in the states that introduce them with changes in the states that do not. But the Grodsky study effectively tosses out all the information available for the 27 states that do not have an exit examination before 2004. As important, the analysis does not consider any measures of state policies except for exit exams, implying that any other policy changes for the three decades between 1971 and 2004 are either irrelevant for student performance or are not correlated with the introduction and use of exit exams.

The central finding is that exit exams do not have a statistically significant effect on test scores. But this insignificance could arise because of any or all of the above-mentioned problems rather than the absence of an effect of exit exams, as the NRC committee wants us to presume.

The committee's estimate of the effects of exit exams on school dropout rates is less controversial. It relies on two quite reliable studies, although they are not without limitations: they study the effects of specific exit exams, which may not generalize to other arrangements. The studies indicate that perhaps 2 percent of potential high-school graduates would have received the diploma had it not been for the exit exams.

The committee touts the possibility of alternative incentives to exit exams: "Several experiments with providing incentives for graduation in the form of rewards, while keeping graduation standards constant, suggest that such

incentives might be used to increase high school completion." The key of course is just what the phrase "while keeping graduation standards constant" means. The idea behind exit exams is to ensure a minimum level of quality, as distinct from meeting the course completion requirements. Moreover, the report never makes the case that exit exams and other potential incentive programs are mutually exclusive. In principle, nobody would argue against employing other incentive programs as long as they were worth the expense and, as the committee says elsewhere, do not introduce perverse incentives of one kind or another.

The Takeaway

The NRC clearly wants to enter into the current debate about the reauthorization of NCLB. And the NRC has an unmistakable opinion: its report concludes that current test-based incentive programs that hold schools and students accountable should be abandoned. The report committee then offers three recommendations: more research, more research, and more research. But if one looks at the evidence and science behind the NRC conclusions, it becomes clear that the nation would be ill advised to give credence to the implications for either NCLB or high-school exit exams that are highlighted in the press release issued along with this report.

The framing of policy in the NRC report is simple: "The small or nonexistent benefits that have been demonstrated to date suggest that incentives need to be carefully designed and combined with other elements of the educational system to be effective." Nobody would oppose careful design of incentives. Nobody would oppose evaluating the intended and unintended outcomes of incentives. And nobody would oppose combining carefully designed incentives with "other elements of the educational system to [make them] effective."

The NRC is careful to offer no guidance on how NCLB or state exit exams might be modified to make them more effective. And the NRC is very careful not to offer any guidance on "other elements of the educational system." The message that comes through is clear: keep working on test development, but never use tests for any incentive or policy purposes.

A better takeaway message might be, "Never rely on the conclusions of this NRC report for any policy purpose."

Eric Hanushek is senior fellow at the Hoover Institution of Stanford University and member of the Koret Task Force on K–12 Education.