Testing, Learning, and Teaching: The Effects of Test-based Accountability on Student Achievement and Instructional Time in Core Academic Subjects

• • •

Martin West
Assistant professor of education, political science, and public policy at Brown University

Is what gets tested what gets taught? Does what gets tested get learned? These questions go to the heart of debates over the national push toward test-based accountability in public education. Proponents argue that the regular testing of students, combined with meaningful accountability for results, will align instruction with curricular standards and motivate students and teachers to work harder. This logic appears to be widely accepted among the American public, who consistently express support for the use of testing as a strategy to improve student achievement.[1]

But test-based accountability is not without its critics. Some prominent education scholars assert that a heavy reliance on testing distorts instruction and undermines authentic learning. What gets tested, goes the argument, gets taught badly—especially if the results have meaningful consequences for schools, educators, or students. Moreover, the introduction of extrinsic rewards and sanctions for students may devalue learning and discourage especially those students who are most in need of improvement. If such criticisms are valid, test-based accountability will fail to improve student achievement even in tested subjects.[2]

A second line of criticism, however, focuses instead on the specific accountability systems currently in use in most states—systems shaped largely by the mandates of the federal No Child Left Behind (NCLB) Act. Enacted in 2002, NCLB requires states to test students annually in grades three through eight (and once in high school) in reading and math. Starting next year, states must also test students in science, but these tests need only be administered in three grades and do not have to be used to determine whether schools are making "adequate yearly progress." Other core subjects, including history, are ignored altogether. Some observers fear that, even if NCLB can be expected to improve reading and math skills, these gains will come at the expense of performance in subjects that go untested and therefore will go untaught.

Concern over the possibility of curricular narrowing under NCLB intensified last March, when a front-page article in the *New York Times* asserted that "thousands of schools across the nation are responding to the reading and math requirements laid out in No Child Left Behind…by reducing class time spent on other subjects and, for some low-proficiency students, eliminating it."[3] Readers of the Center on Education Policy (CEP) study cited in the article may have been surprised to find that only one-third of the district officials surveyed during the 2005–06 school year reported that elementary schools had reduced time spent on social studies or science either "somewhat" or "to a great extent," while just over one-fifth said the same about art and music.[4] As Craig Jerald pointed out in a Center for Comprehensive School Reform and Improvement policy brief, the study suggested an "erosion—rather than a decimation—of subjects outside of reading and mathematics."[5] Yet the story seemed to tap into widespread fears about the consequences of a national accountability system based solely on a few subjects. Media reports of schools abandoning art, music, and physical education to focus on reading and math proliferated. In June, renowned historian David McCullough took his concerns to Congress, testifying that "because of the No Child Left Behind program, sadly, history is being put on the back burner or taken off the stove altogether in many schools, if not most schools, with the argument that we have to concentrate on reading and mathematics and science."[6]

McCullough's account of a narrowing curriculum sounds plausible, but is it true? If so, should policy makers abandon the entire accountability enterprise? Or is NCLB worth saving—perhaps by mandating tests in additional subjects?

Unfortunately, solid data to address these questions are scarce. Although the proliferation of state and federal testing programs has yielded an abundance of information about what American students know, our knowledge of what they are taught remains fragmentary and incomplete. Below, I present new information on trends between 1988 and 2004 in the amount of time elementary school teachers nationwide spent on instruction in each of four core academic subjects. The information, which is based on teacher self-reports, provides new insights about the extent of curricular narrowing during the initial implementation of NCLB and the degree to which state testing policies influence the allocation of instructional time.

First, however, I review what is known about the effects of introducing statewide testing systems on student achievement in the subjects that have been the focus of state and federal accountability efforts to date: reading and math.

Testing and Learning

One claim about the effects of test-based accountability appears beyond dispute. As testing experts Laura Hamilton and Brian Strecher of RAND put it, "the implementation of high-stakes testing has almost always led to increases in test scores."[7] Indeed, a large collection of literature indicates that the introduction of test-based accountability systems is typically associated with a sharp increase in reported student performance on the specific assessments used.[8] Yet the extent to which these increases are meaningful remains a matter of debate. This is because a wide variety of factors can lift test scores in the absence of authentic student learning, a phenomenon known as test-score inflation.[9]

The most straightforward causes of test-score inflation are coaching or outright cheating on the part of teachers. It is important to distinguish inappropriate coaching from the practice of "teaching to the test." Although accountability critics denigrate teaching to the test, proponents see the alignment of instruction with well-designed assessments as one of accountability's main benefits. However, when teachers are familiar with the tests their students will take and adjust their lessons to cover specific test items, as opposed to covering the broader domain from which those items are drawn, test scores become unreliable indicators of student knowledge.[10] The problem is even more obvious when teachers directly assist students during the administration of a test or answer questions on their behalf.[11] Inappropriate coaching and cheating have relatively simple solutions—namely, rotating test items regularly, safeguarding test materials, and adopting sensible test-administration practices—but the extent to which they are responsible for observed increases in test scores after the introduction of accountability systems is often difficult to assess.

Moreover, a host of more sophisticated strategies are available to schools that face strong incentives to demonstrate rapid progress toward higher test scores. For example, schools may classify more students as needing special education, retain more students at grade level, or issue suspensions to coincide with test dates to alter the population taking the test.[12] Some schools in Virginia have even been shown to serve high-calorie meals on the days when state exams are administered, presumably in an effort to enhance alertness and concentration.[13] Such strategies are often only feasible in the short run and can be circumvented by requirements that a specific percentage of students be tested each year. Their availability, however, again calls into question the usefulness of publicly reported data on test performance to gauge the effects of existing accountability systems.

The likelihood of test-score inflation places a premium on sources of student performance data that are independent of any given high-stakes accountability sys-

tem. Probably the best such evidence currently available is the National Assessment of Educational Progress (NAEP), which provides state-level measures of student achievement in fourth and eighth grade at regular intervals. Because the NAEP is a "no-stakes" exam—the scores of individual districts, schools, and students are not even reported—there is scant incentive to cheat or even to prepare students for the test, nor to manipulate the test-taking population to inflate scores artificially.[14]

To my knowledge, four published studies have used state NAEP data to evaluate the effects of test-based accountability policies on student achievement. It is important to note that none of the four studies examines the effects of NCLB. In fact, the implementation of NCLB has further complicated the task of measuring accountability's effects: With all states now engaged in a broadly similar set of policies, cross-state comparisons have been rendered all but useless. Instead, each of the studies attempts to gauge the effects of test-based accountability by examining NAEP performance in the 1990s, when many states were in the process of implementing their own accountability systems. The number of states with accountability systems in place increased from 4 in 1993 to 40 by 2000.[15] Although participation in the state NAEP was voluntary before the passage of NCLB, roughly 40 states took part in assessments of math performance in 1992, 1996, and 2000, and of reading performance in 1994, 1998, and 2002.

One early study, by Audrey Amrein and David Berliner of Arizona State University, set out to compare changes in NAEP reading and math scores following the implementation of "high-stakes" accountability systems in 28 states with concurrent trends among all states participating in the NAEP.[16] Because the authors eliminated states where exclusion rates on the NAEP moved in the same direction as test scores (e.g., an increase in both exclusion rates and test scores), however, the number of states for which they reported results was actually considerably smaller. When their analysis showed more high-stakes states losing ground against the national average than gaining ground, the authors deemed test-based accountability a "failed policy initiative" and argued that it was time for policy makers to move on.

Although the Amrein and Berliner study received considerable media attention, it was fundamentally flawed. Most important, the authors compared each high-stakes testing state individually to the nation as a whole during a period in which many other states were in the process of adopting test-based accountability policies. As Eric Hanushek and Margaret Raymond of Stanford University soon showed, simply comparing the high-stakes testing states as a group with states without high-stakes testing, even while maintaining the same classification scheme Amrein and Berliner used,

revealed a positive relationship between accountability and student achievement.[17]

The other three studies of the effects of test-based accountability during the 1990s differ in their details, but they all yield more favorable findings. Hanushek and Raymond examined the rate of progress in math over two four-year spans, one starting in 1992 and one in 1996, and compared states with and without accountability systems in place during these two time periods.[18] Martin Carnoy and Susannah Loeb, also of Stanford, created a 1–5 index measuring the extent of accountability pressure in each state to determine whether the degree of accountability in a state affected the changes in fourth- and eighth-grade math scores between 1996 and 2000.[19] Most recently, Hanushek and Raymond updated their analysis to include results from the 2002 NAEP reading test. These additional data allowed them to measure the progress of states that adopted accountability systems against their own performance earlier in the decade, offering a particularly rigorous test of accountability's effects.[20] All three studies indicate that states that adopted accountability systems during this period improved their performance relative to states without such systems.

Taken together, these results provide convincing evidence that the adoption of test-based accountability policies in the 1990s contributed to the overall increases in fourth- and eighth-grade math scores on the NAEP during this period. Between 1990 and 2003, NAEP math scores increased by 25 points for fourth graders and by 16 points for eighth graders, historically unprecedented gains equivalent to roughly two full years of student learning. Although overall gains on the NAEP reading test in the 1990s were quite small, Hanushek and Raymond's most recent analysis indicates that reading performance, too, was enhanced by the adoption of accountability systems.[21]

At the same time, it is important to note several caveats associated with this research. First, the beneficial effects of accountability reported in each study were all modest in size, equivalent to about 0.2 standard deviations in student achievement. Because the NAEP is not perfectly aligned with state standards, this figure likely understates the true effects of accountability on performance on state tests.[22] Even so, the gains fall far short of what will be necessary to meet the goal of universal proficiency established by NCLB. Nor does it appear that accountability systems have had a larger impact on the performance of minority or low-income students, suggesting that additional strategies will be necessary to close achievement gaps.[23] Finally, although Hanushek and Raymond find that state testing systems need to include consequences for underperformance to work (because systems that simply required the public reporting of test results did not result in higher achievement), these studies

provide little indication of which types of accountability provisions are most effective.

The modest effects of test-based accountability to date may reflect the relatively weak incentives that have been used in most state systems and, in particular, the absence of strong incentives for individual students and teachers. The focus of existing accountability systems has been on schools, which may not be the best target for generating a behavioral response. Accountability systems that create strong incentives for individual students by requiring them to pass state tests to advance to the next grade, for instance, have generated larger gains in student achievement.[24] And the presence of centrally administered, curriculum-based exit exams with meaningful consequences for individual students has consistently been shown to be a key factor in explaining international differences in student achievement.[25] Evidence on the effect of performance incentives for individual teachers is more limited, but it suggests that merit- or performance-based pay systems may be effective in improving student achievement.[26]

In sum, research suggests that what gets tested does in fact get learned, at least when testing is combined with some degree of accountability for results. Moreover, there is no credible evidence that testing reduces achievement in tested subjects. Although these results do not bear directly on the effectiveness of NCLB, they provide a strong rationale for the use of test-based accountability as one component of our national effort to improve student achievement.

Testing and Teaching

Yet the concern remains that the heightened focus under NCLB on student achievement in reading and math (and, to a lesser extent, science), however important these subjects may be, may distract schools from other important goals. The correct balance between instruction in basic skills and in other subjects is a matter of debate, especially in the early grades. If, however, as the now-defunct Council on Basic Education (CBE) argued a few years back, "Every American child deserves [an education which] comprises challenging, standards-based instruction in English, mathematics, history, civics, geography, foreign-language, and the arts," the possibility that schools are reducing or eliminating altogether instruction in untested subjects warrants serious attention.[27]

As noted above, there is suggestive evidence that some curricular narrowing has already occurred since NCLB was passed. In addition to the CEP survey discussed above, about three-quarters of respondents to a CBE survey of principals in Illinois, Maryland,

New York, and New Mexico conducted during the 2003–04 school year reported having increased instructional time in reading, writing, and math since 2000. Among elementary school principals, more than one-quarter reported decreases in instructional time in social studies, civics, and geography, including 47 percent of principals in high-minority schools. The latter figure raises the disturbing possibility that curricular narrowing may be most severe in schools with heavily minority student bodies, for whom the pressure to raise achievement in the reading and math is most severe.[28]

Unfortunately, so far little systematic evidence has been available on the amount of instruction actually delivered in core academic subjects, information that is essential to determine the extent to which administrator surveys and anecdotal reports accurately portray the experiences of most American students. Nor is it possible from these surveys to gauge the magnitude of any changes in curricular emphasis.

To remedy this gap, Table 1 presents data from the U.S. Department of Education's Schools and Staffing Survey (SASS) on the amount of time that first- and sixth-grade teachers reported spending each week in four core academic subjects: English, math, history or social studies, and science.[29] The data cover the period from 1987–88 until 2003–04, the second school year after the enactment of NCLB. The analysis is limited to full-time teachers in self-contained classrooms to ensure that their responses capture the classroom experiences of individual students. The percentage of elementary school teachers who teach in self-contained classrooms increased from just under 60 percent in 1989–90 to roughly 75 percent in 1993–94, but it has remained roughly constant since that time.

The data confirm a marked increase in the amount of instruction elementary school students received in reading during the initial implementation of NCLB. Weekly time spent on reading instruction increased by roughly 40 minutes between the 1999–00 and 2003–04 school years. This increase did not correspond with an overall increase in the total amount of time spent on instruction in core academic subjects, which actually declined slightly over the same four years. Rather, teachers' reports indicate a modest decrease in time spent on instruction in each of the other three subjects, including math. Weekly instructional time fell by 17 minutes in math, by 23 minutes in history, and by 17 minutes in science. As a percentage of the time spent in each subject in 1999–00, the declines were 5 percent, 13 percent, and 10 percent, respectively.

TABLE 1. Weekly Instructional Time in Core Academic Subjects, First through Sixth Grades, 1987–2004

	1987–88 *(hours)*	1989–90 *(hours)*	1993–94 *(hours)*	1999–00 *(hours)*	2003–04 *(hours)*	Change, 1988–2004 *(minutes)*	Change, 1999–2004 *(minutes)*
Reading/English Language Arts	10.7	10.2	10.6	10.6	11.3	+36.6	+39.6
Mathematics	4.9	4.9	5.2	5.7	5.3	+28.8	-17.4
Science	2.8	2.8	3.1	2.7	2.5	-17.4	-22.8
History/ Social Studies	3.0	3.0	3.2	3.0	2.6	-21.6	-16.8
Total *(Core Subjects)*	21.2	20.8	22.1	22.0	21.7	+29.4	-16.8
N	**9,824**	**9,499**	**8,376**	**7,244**	**7,397**	**—**	**—**
Other Activities	10.3	11.0	10.1	10.6	10.9	+31.8	+16.8
School Week	31.6	31.9	32.1	32.6	32.6	+61.2	0
First- to sixth-grade teachers in self-contained classrooms (%)	58.9	58.6	75.7	74.4	78.1	—	—
N	**9,270**	**9,190**	**7,784**	**6,675**	**6,919**	**—**	**—**

Source: Schools and Staffing Survey, U.S. Department of Education, various years.
Note: The sample size is smaller for the length of the school week because of district-level nonresponse.
— = not applicable.

Of course, fears about tests eroding instruction in subjects other than reading and math predate the enactment of NCLB. As early as 1991, Lorrie Shepard of the Center for Research on Evaluation, Standards, and Student Testing wrote that "Although critics may originally have feared that testing would take instructional time away from 'frills,' such as art and citizenship, the evidence now shows that social studies and science are neglected because of the importance of raising test scores in the basic skills."[30] The SASS data indicate that the amount of instruction elementary school students nationwide received in both history and science actually increased slightly during the first half of the 1990s. After 1993–94, however, as more states implemented accountability systems, time spent on instruction in these two subjects began to decline, while time spent on math

increased sharply. Using 1987–88 as a baseline, instructional time in English and math had increased by 37 minutes and 29 minutes, respectively, by 2003–04. Time spent on history declined by 22 minutes and time spent on science declined by 17 minutes. Table 2 tracks these fluctuations over this period in the relative amount of time spent on each of these four subjects, revealing that the share of instruction in core academic subjects devoted to reading and math has increased from 73 percent to 77 percent.

TABLE 2. Instructional Time as a Percentage of Time Spent on Instruction in Core Academic Subjects, First through Sixth Grades, 1987–2004

	1987–88 *(percent)*	1989–90 *(percent)*	1993–94 *(percent)*	1999–2000 *(percent)*	2003–04 *(percent)*	Change, 1988–2004 *(percent)*
Reading/English Language Arts	50.2	48.8	47.8	48.3	51.9	+1.7
Mathematics	23.0	23.4	23.6	25.7	24.7	+1.7
Science	13.0	13.3	14.1	12.5	11.3	-1.6
History/Social Studies	14.0	14.5	14.4	13.6	12.1	-2.0
Total	100.0	100.0	100.0	100.0	100.0	—
N	**9,824**	**9,499**	**8,376**	**7,244**	**7,397**	**—**

Source: Schools and Staffing Survey, U.S. Department of Education,various years.
Note: Columns may not sum to 100 because of rounding. — = not applicable.

Although they are the basic building blocks of almost any solid elementary-school curriculum, English, math, science, and history do not encompass the full range of experiences and knowledge that students will gain in elementary school. The arts, music, physical education, and even recess are all crucial for the development of healthy, culturally literate citizens. The SASS unfortunately does not contain useful data for directly gauging the extent of students' exposure to these other aspects of a liberal arts curriculum. However, we can get a rough sense of the amount of time available for enrichment activities by examining trends in the total amount of time students spend at school and in the time devoted to instruction in the four core subjects for which data are available.

Table 1 shows that the total amount of time students spent in school each week increased by a full hour between 1987–88 and 2003–04, with half of this increase taking place by 1993–94 and no change evident following the implementation of NCLB.

Meanwhile, the total amount of time spent on instruction in core subjects increased by less than half an hour. Assuming that the time taken up by routine activities such as dismissal, lunch, and assemblies remained constant, the result has been a modest increase in the total amount of time available for enrichment activities. This finding is consistent with the CEP surveys of administrators, which show that physical education has been less affected by the implementation of NCLB than has instruction in other academic subjects.[31]

In sum, the initial implementation of NCLB was accompanied by a substantial increase in the amount of instruction that elementary school students received in reading and by modest declines in the time spent on math, science, and history. During the late 1990s, when many states adopted accountability systems, instructional time increased sharply in math and declined in history and science. Although these descriptive data cannot definitively establish a causal relationship between testing policies and instructional time, the patterns are consistent with an accountability-driven shift in elementary school curricula toward basic skills in reading and math and away from science and history. The extensive case-study evidence indicating that teachers respond to the introduction of accountability systems in part by increasing the amount of time spent on tested content and subjects makes such an interpretation highly plausible.[32]

At the same time, the changes in time use in elementary education in recent years are less pronounced than the rhetoric surrounding the issue of curricular narrowing would suggest. This does not invalidate accounts of specific schools in which instruction in subjects other than reading and math has been sharply decreased or eliminated altogether. Nor does it mean that the effects of NCLB on curricula will not become more pronounced over time, as performance targets increase and more schools are identified as needing improvement. But it does suggest that NCLB's impact on elementary schools' curricula, at least through 2003–04, was gentler than has been thought.

Testing and Instructional Time in Science and History

What steps should state and federal policy makers wanting to guard against curricular narrowing consider? One possibility is to introduce additional tests in history and perhaps incorporate performance on those tests into the definition of adequate yearly progress. Testing researcher Richard Phelps makes this case forcefully: "If not-tested subjects are being dropped, either they, too, should be tested or, perhaps, educators and policy makers are signaling that, in a world of tough choices among competing priorities, some subjects must in fact take a backseat to others."[33] Whether

intentional or not, the signal now being sent by NCLB clearly prioritizes the development of basic skills in reading and math over other important subject areas.

A closer look at the 2003–04 SASS data suggests that introducing testing in science and history is likely to be effective in increasing the time devoted to instruction in those subjects. Table 3 presents the average time spent on science and history for states that do and do not have elementary school assessments in those subjects. Although NCLB does not require testing in history, and will not require testing in science until the 2007–08 school year, by 2003–04, 23 states had tested elementary students in at least one grade in science using assessments aligned to state content standards, and 16 states had done so in history or social studies.[34] Few of these states included performance in either subject to determine whether schools made adequate yearly progress under NCLB or in state accountability ratings, and many tested elementary school students in science or history only in a single grade, presumably making a strong relationship between policies and patterns of time use less likely.

TABLE 3. Weekly Instructional Time in Science and History, First through Sixth Grades, 2003–04

	All Schools *(hours)*	Testing in Subject *(hours)*	No Testing in Subject *(hours)*	Difference *(minutes)*
Science	2.46	2.74	2.17	34.2
History	2.62	2.88	2.47	24.6
Total for Core Subjects	21.7	21.8	21.6	12.0
N	**7,397**	**3,115 (science)** **2,202 (history)**	**4,282 (science)** **5,195 (history)**	**—**

Source: Schools and Staffing Survey, U.S. Department of Education, 2004; Education Week, Quality Counts 2004 (Bethesda, MD: Education Week Press).
Note: Testing in subject indicates that the school is located in a state with standards-aligned assessments in science or history in at least one elementary school grade. — = not applicable.

Even so, Table 3 reveals that the differences between the time spent on science and history in states that do and do not test in those subjects are immense—on the order of 34 minutes in science and 25 minutes in history. That's 26 percent more instructional time for states that test science, and 17 percent more instructional time for states that test history. These differences far exceed the observed decline in the amount of time spent on science and history instruction since the adoption of

NCLB. As appendix table A1 shows, these differences persist after adjusting the data for differences in the percentage of a school's students who are minorities, the percentage eligible for the free and reduced-price lunch program, whether the school is located in an urban environment, and the teacher's experience.

As an additional point of comparison, Table 4 presents similar information for schools with high-minority (80 percent or greater) and low-minority (20 percent of fewer) student bodies. As noted above, the possibility that the impact of NCLB on curricular breadth has been especially severe for minority students has been widely discussed. The results show that high-minority schools do spend less time on science and history than low-minority schools, despite the fact that they spend more total time each week on core academic subjects. The differences, however, are minute: about five weekly minutes in each subject.

TABLE 4. Weekly Instructional Time in Science and History, First through Sixth Grades, 2003–04

	All Schools *(hours)*	Low-Minority Schools *(hours)*	No High-Minority Schools *(hours)*	Difference *(minutes)*
Science	2.46	2.51	2.42	5.4
History	2.62	2.68	2.60	4.8
Total for Core Subjects	21.7	21.3	22.0	42
N	**7,397**	**1,676**	**3,224**	**—**

Source: Schools and Staffing Survey, U.S. Department of Education, 2004.
Note: High-minority schools are schools with at least 80 percent minority student bodies; low-minority schools are schools with fewer than 20 percent minority student bodies. — = not applicable.

Simple differences in instructional time across states cannot establish that the presence of testing in a particular subject causes teachers to spend more time on those subjects. It is possible—even likely—that states that have adopted testing policies in science or history in the absence of a federal mandate are more committed to educating students in those subjects for other reasons. Yet the success of states that are testing students in science and history in sustaining instruction in those subjects in the NCLB era indicates that this may be a worthwhile step for Congress to consider as the law's reauthorization date approaches.

Conclusion

Discussions of curricular narrowing as a result of NCLB have taken place in an empirical vacuum, which the instructional time data presented here can only incompletely fill. More important, while teachers may report that they are spending 2.6 hours each week on history or social studies, we still have little idea of how well those hours are being spent. Independent evaluations of standards in history and science find that the quality of these standards vary widely across states.[35] These differences in quality are probably at least as important an influence on students' academic progress as differences in the quantity of time spent covering a particular subject area. Put differently, simply increasing the time spent on ineffective instructional practices is unlikely to improve student achievement. Although test-based accountability can be a powerful tool in aligning instruction to content standards, careful attention needs to be paid to the standards themselves.

Moreover, measuring the amount of time teachers devote to reading, math, science, and history ignores the extent to which instruction in multiple subjects can and should proceed simultaneously. As McCullough told Congress, it's "fine, to concentrate on the reading all they want. But they don't just have to read what is conventionally seen as literature. They can read the literature of history."[36] The opportunities for synthesizing instruction in math and science are equally clear. Especially for elementary school students, the allocation of time among different subjects is not a zero-sum game.

Thanks largely to the work of E. D. Hirsch, we now better comprehend the critical role that a curriculum rich in content knowledge across a wide range of subjects plays in the development of reading ability. To the extent that becoming a strong reader—one capable of passing state proficiency standards—requires strong content knowledge in areas such as history, science, and even the arts, the narrowing of the curriculum in response to accountability pressures in reading and math may be a self-correcting problem. If intensive instruction in decoding ultimately fails to yield desired improvements in reading proficiency as students age, schools may recognize the shortcomings of a curriculum based solely on those skills. In the meantime, introducing high-quality tests in science and especially in history may help to ensure that the modest narrowing of the curriculum that has occurred to date does not accelerate.

TABLE 5. The Relationship between State Testing Policies, School Characteristics, and Time Spent on Science and History, 2003–04

	(1) Science	(2) Science	(3) History	(4) History
Elem Test in Science	0.582***	0.580***	—	—
	(0.163)	(0.157)		
Elem Test in History	—	—	0.401***	0.405***
			(0.127)	(0.127)
Percent Minority	—	-0.0001	—	0.001*
		(0.002)		(0.0007)
Percent Free Lunch	—	-0.0005		-0.002**
		(0.002)	—	(0.0009)
Urban	—	-0.134		-0.110
		(0.086)	—	(0.101)
Teacher Experience	—	-0.006		0.0008
		(0.004)	—	(0.003)
Second Grade	-0.013	-0.012	-0.026***	-0.027***
	(0.008)	(0.009)	(0.008)	(0.008)
Third Grade	-0.005	-0.005	-0.024***	-0.024***
	(0.008)	(0.007)	(0.008)	(0.008)
Fourth Grade	-0.003	-0.004	0.023*	0.023*
	(0.009)	(0.009)	(0.013)	(0.012)
Fifth Grade	0.040***	0.040***	0.040***	0.040***
	(0.008)	(0.008)	(0.010)	(0.010)
Sixth Grade	0.088**	0.087**	0.062***	0.062***
	(0.035)	(0.034)	(0.012)	(0.012)
Constant	2.946***	3.069***	3.040***	3.071***
	(0.352)	(0.336)	(0.204)	(0.210)
R2	**0.04**	**0.04**	**0.04**	**0.04**
N	**7,397**	**7,397**	**7,397**	**7,397**

Source: Author's calculation using the Schools and Staffing Survey, U.S. Department of Education, 2004. Information on state testing policies is drawn from Education Week, Quality Counts 2004 *(Bethesda, MD: Education Week Press, 2004). Note: Ordinary least squares regression; standard errors adjusted for clustering by state in parentheses. States identified as testing in science and history are those that tested elementary students in at least one grade in those subjects using assessments aligned to state content standards during the 2003-04 school year. * = p <0.1; ** = p <0.05; *** = p <0.01; — = not applicable.*

Endnotes

[1] Richard P. Phelps, "Persistently Positive: Forty Years of Public Opinion on Standardized Testing," in *Defending Standardized Testing,* ed., Richard P. Phelps (Mahwah, NJ: Lawrence Erlbaum Associates, 2005), 1–23; Richard P. Phelps, "The Demand for Standardized Student Testing," *Educational Measurement and Practice* 17, no. 3 (1998): 5–23.

[2] See, for example, Linda Darling Hammond, "Standards, Accountability, and School Reform," *Teachers College Record* 106, no. 6 (June 2004): 1047–1085; Alfie Kohn, *The Case Against Standardized Testing* (Portsmouth, NH:Heinemann, 2000); George F. Madaus, "The Distortion of Teaching and Learning: High-Stakes Testing and Instruction," *Peabody Journal of Education* 65, no. 3 (Spring 1988): 29–46; and Nell Noddings, "Care and Coercion in School Reform," *Journal of Educational Change* 2, no. 1 (2001): 35–43.

[3] Sam Dillon, "Schools Cut Back Subjects to Push Reading and Math," *New York Times*, March 26, 2006, A1.

[4] Center on Education Policy, *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act* (Washington, DC: Center on Education Policy, 2006).

[5] Craig G. Jerald, "The Hidden Costs of Curriculum Narrowing," Center for School Reform and Improvement, August 2006, available at http://www.centerforcsri.org/files/CenterIssueBriefAug06.pdf.

[6] Testimony of David McCullough, Hearing before the Subcommittee on Education and Early Childhood Development of the Committee on Health, Education, Labor, and Pensions, United States Senate, 109th Congress, 1st Session, June 30, 2005.

[7] Laura S. Hamilton and Brian M. Strecher, "Improving Test-Based Accountability," in *Making Sense of Test-Based Accountability in Education,* eds. Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, CA: Rand, 2002), 121.

[8] Richard P. Phelps, "The Rich, Robust Research Literature on Testing's Achievement Benefits," in *Defending Standardized Testing,* ed. Richard P. Phelps (Mahwah, NJ **:** Lawrence Erlbaum Associates, 2005), 55–90.

[9] The tests administered as part of most accountability systems consist of samples of questions drawn from the broader domain of knowledge that the test is designed to assess. The test scores of individual students measure, inevitably with some degree of imprecision, the students' mastery of that broader domain of knowledge. When test scores are inflated, they no longer support valid inferences about the subject matter being assessed. For an excellent discussion of the meaningfulness of test score gains, see Brian A. Jacob, "A Closer Look at Achievement Gains under High-Stakes Testing in Chicago," in *No Child Left Behind? The Politics and Practice of School Accountability,* eds. Paul E. Peterson and Martin. R. West (Washington, DC: Brookings Institution Press, 2003), 269–91.

[10] Brian M. Strecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education,* eds. Laura S. Hamilton, Brian M. Strecher, Stephen P. Klein (Santa Monica, CA: Rand, 2002), 79–100.

[11] Brian A. Jacob and Steven Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118, no. 3 (2003): 843–77.

[12] David N. Figlio and Lawrence S. Getzler, "Accountability, Ability, and Disability: Gaming the System," NBER Working Paper No. 9307, National Bureau of Economic Research, October 2002; David N. Figlio, "Testing, Crime, and Punishment," NBER Working Paper No. 11194, National Bureau of Economic Research, March 2005; Julie B. Cullen and Randall Reback, "Tinkering Toward Accolades: School Gaming Under a Performance-Based Accountability System," Working Paper, University of Michigan, 2002.

[13] David N. Figlio and Joshua Winicki, "The Effects of School Accountability Plans on School Nutrition," *Journal of Public Economics* 89, no. 2–3 (February 2005): 381–94.

[14] Although exclusion rates on the NAEP do appear to be systematically related to performance (states that exclude more special education and limited English proficiency students tend to see their test scores increase), data on the share of students that is excluded are readily available to make statistical adjustments that account for this bias. Eric A. Hanushek and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24, no. 2 (2005): 297–327.

[15] Hanushek and Raymond, "Does School Accountability Lead to Improved Student Performance?" 297–327.

[16] Audrey L. Amrein and David C. Berliner, "High-Stakes Testing, Uncertainty, and Student Learning," *Education Policy Analysis Archives* 10, no. 18 (March 2002), available at http://epaa.asu.edu/epaa/v10n18/ (accessed April 23, 2007).

[17] The additional gains made by states with high-stakes testing systems remained statistically significant even after adjusting for changes in NAEP exclusion rates. Eric A. Hanushek and Margaret E. Raymond, "High-Stakes Research," *Education Next* 3, no. 3 (Summer 2003): 48–55; for a similar critique, see Barak Rosenshine, "High-Stakes Testing: Another Analysis," *Education Policy Analysis Archives* 11, no. 24 (August 2003), available at http://epaa.asu.edu/epaa/v11n24/ (accessed April 23, 2007).

[18] Eric A. Hanushek and Margaret E. Raymond, "Improving Educational Quality: How Best to Evaluate our Schools?" in *Education in the Twenty-First Century: Meeting the Challenges of a Changing World,* ed. Yolanda Kodrzycki (Federal Reserve Bank of Boston, 2002), 193–224.

[19] Martin Carnoy and Susannah Loeb, "Does External Accountability Affect Student Outcomes?" *Educational Evaluation and Policy Analysis* 24, no. 4 (Winter 2002): 305–31.

[20] Hanushek and Raymond, "Does School Accountability Lead to Improved Student Performance?" 297–327.

[21] Tom Loveless, "How Well Are American Students Learning?" 2006 Brown Center Report on American Education (Washington, DC: Brookings Institution, 2006).

[22] Improvements in state performance on the NAEP are generally smaller than improvements on state tests. The extent to which the differences reflect artificial inflation of state test scores or simply the lack of alignment between state and NAEP standards is difficult to assess. See Daniel M. Koretz and Sheila I Barron, *The Validity of Gains on the Kentucky Instructional Results Information System* (Santa Monica, CA: Rand Corporation, 1998); Stephen P. Klein, Laura S. Hamilton, Daniel F. McCaffrey, and Brian M. Stecher, "What do Test Scores in Texas Tell Us?" Issue Paper No. 202, Rand Corporation, 2000.

[23] Hanushek and Raymond, "Does School Accountability Lead to Improved Student Performance?" 297–327.

[24] Brian A. Jacob, "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89, no. 5–6 (2005): 761–96.

[25] Ludger Woessmann, "Central Exams and Student Achievement: International Evidence," in *No Child Left Behind? The Politics and Practice of School Accountability,* eds. Paul E. Peterson and Martin. R. West (Washington, DC: Brookings Institution Press, 2003), 292–323; John H. Bishop, "The Effect of National Standards and Curriculum-Based Exams on Achievement," *American Economic Review* 87, no. 2 (1997): 260–64.

[26] Michael J. Podgursky and Matthew G. Springer, "Teacher Performance Pay: A Review," Working Paper, National Center on Performance Incentives, October 24, 2006.

[27] Claus von Zastrow with Helen Janc, *Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools* (Washington DC: Council for Basic Education, 2004), 11.

[28] von Zastrow with Janc, *Academic Atrophy,* 7-8.

[29] For an earlier analysis of instructional time using data from the Schools and Staffing Survey, see Marianne Perie, David P. Baker, and Sharon Bobbitt, "Time Spent Teaching Core Academic Subjects in Elementary Schools: Comparisons Across Community, School, Teacher, and Student Characteristics." U.S. Department of Education, National Center for Education Statistics (NCES 97-293), 1997.

[30] Lorrie A. Shepard, "Will National Tests Improve Student Learning?" *Phi Delta Kappan* 72, no. 1 (November 1991): 233–34.

[31] Center on Education Policy, *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act* (Washington, DC: Center on Education Policy, 2006).

[32] Strecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," 79–100.

[33] Richard P. Phelps, "Why Testing Experts Hate Testing," in *Testing Student Learning, Evaluating Teacher Effectiveness,* eds. Williamson M. Evers and Herbert J. Walberg (Stanford, CA: Hoover Institution Press, 2004).

[34] Information on state testing policies is drawn from Education Week, *Quality Counts 2004* (Bethesda, MD: Education Week Press, 2004).

[35] Chester E. Finn, Jr., Michael J. Petrilli, and Liam Julian, *The State of State Standards: 2006* (Washington, DC: Fordham Foundation, 2006).

[36] Testimony of David McCullough, Hearing before the Subcommittee on Education and Early Childhood Development of the Committee on Health, Education, Labor, and Pensions, United States Senate, 109th Congress, 1st Session, June 30, 2005.