

Performance-Based Accountability Policies: Implications for School and Classroom Practices
Jane Hannaway, The Urban Institute
Laura Hamilton, RAND

**PERFORMANCE-BASED ACCOUNTABILITY POLICIES:
IMPLICATIONS FOR SCHOOL AND CLASSROOM PRACTICES**

Jane Hannaway
The Urban Institute
JHannawa@ui.urban.org

and

Laura Hamilton
RAND
laurah@RAND.org

This paper focuses on school- and classroom-level responses to performance-based accountability. While the ultimate outcome of interest of education accountability policies is student achievement, a focus on intermediary outcomes – such as school and classroom behavior – has important policy implications in its own right for reasons discussed below. This paper is one of six papers prepared for the U. S. Department of Education in response to Section 1503 of the No Child Left Behind legislation requiring that the U. S. Department of Education provide an interim report to Congress on the effects of the law.

BACKGROUND

The last two decades have witnessed a marked shift in education policy from a focus on changes in inputs and process to a focus on outcomes – in particular, student achievement as measured by test performance. Key states, including Texas, North Carolina, Kentucky, Maryland, and Florida, broke the path in the late 1980s and 1990s by holding schools accountable for the learning of their students. In 2001 the policy went national. Congress passed the landmark No Child Left Behind Act (NCLB) requiring that schools and school districts across the country be held accountable for student performance.¹ States must set targets for school and district performance and assess whether schools and districts make adequate yearly progress (AYP) toward a goal of proficiency for all students. Schools and districts that do not make adequate progress are subject to interventions.

¹ Earlier authorizations of the Elementary and Secondary Education Act (ESEA) required standards and tests, but did not enforce accountability mechanisms.

The ideas behind NCLB are simple ones. The intent is to promote improved student achievement and to reduce achievement gaps among student groups -- especially gaps between advantaged and disadvantaged students -- through the establishment of standards and performance-based accountability mechanisms. Standards define important areas of student learning by subject and grade level; performance-based accountability mechanisms establish incentives for teachers and schools to promote learning in those areas. Many accountability systems also include provisions for assistance that is provided to schools or districts that do not meet their accountability targets. The alignment of curriculum and instructional practices with the standards is a critical implied part of the reform and the focus of this paper. Educators are expected to respond to the incentives and the assistance in ways that lead to increased alignment between the curriculum and instruction offered to students and the standards and assessments that reflect the system's learning goals.

Examining school and classroom behavior is important for sorting out the costs and benefits of performance-based accountability systems. Outcome measures alone, namely student test scores, are insufficient for such a determination. They can only tell part of the story largely because no test, or even set of tests, can examine everything we want students to learn. Some important aspects of learning, such as creativity and higher order problem solving, are not very amenable to large-scale paper and pencil testing. In addition, the items included in single test address only a sample of the full range of skills and knowledge that are embodied in state standards, and there are important school

outcomes—both academic and non-academic—that are not tested at all. Because the costs of testing all important areas of learning are exceedingly high, policymakers must rely on other sources of information such as evidence regarding teachers’ behaviors to obtain a complete picture of what outcomes schools are promoting and whether performance-based accountability has led to beneficial or harmful changes to the educational opportunities schools provide to students. Examining behavioral responses is also important because, as we explain later, they can provide early indicators of whether, why and where reforms may be off track.

INTERMEDIATE OUTCOMES

Intermediate outcomes refer here to the behavioral responses of teachers and schools to standards and accountability policies. We are especially interested in classroom instructional practices since this is where student learning takes place. But we are also interested in the system’s response to accountability pressure, for example, its influence on companion policies, such as those that might affect and resource allocation patterns. We report evidence to address four specific questions. The first three questions ask about the extent to which standards and accountability policies shape instructional practice in classrooms. The fourth asks about their influence on the infrastructure that supports instruction.

1. *Do standards and accountability policies lead teachers to focus on different subject areas than they otherwise would?* For example, are teachers spending

more time on tested subjects, most often mathematics and reading? To what extent is this at the expense of other subjects or activities?

2. *Do standards and accountability policies lead teachers to focus on some types of instructional practices or content areas over others?* For example, is there a greater instructional emphasis on the types of skills tested, e.g., basic skills? problem solving skills? inquiry skills?
3. *To what extent do standards and accountability policies lead teachers to focus on skills specific to assessment and testing procedures or to engage in other behaviors that might lead to higher test scores, but have little impact on real learning by students?*
4. *To what extent do standards and accountability policies influence the infrastructure that supports instruction?* For example, are resources reallocated to support teachers in developing the instructional skills and practices necessary to teach to the standards?

Answers to these questions provide insights into the effects of performance-based accountability where direct outcome measures are not readily or fully available either because they do not exist or because they are evident only with some lag. They are important for two major reasons.

First, while a performance-based accountability system provides performance incentives for teachers and administrators by meting out rewards and penalties on the basis of how well students measure up on tests, they also create an inherent problem. To the extent

that some outcomes are more easily measured than others, an accountability system may lead to behavioral distortions (Milgrom and Roberts, 1991; Hannaway, 1992; Ladd, 1996). In short, individuals may be overly attentive to what is measured and slight other important areas of learning. Evidence on what is actually occurring in schools and classrooms can help detect such distortion and gauge the likely magnitude. Without such information, policy makers have at best incomplete information on policy effects.

Second, as suggested earlier, information on intermediate outcomes helps to explain *why* outcomes occur. Such information has three benefits. First, it identifies the mechanisms that produce outcomes and therefore can guide administrators to practices that may more reliably yield hoped-for results. It can also provide insight into behavioral responses that might lead to higher student test scores, but does so in ways that do not lead to real learning, e.g., through teaching test taking skills. In short, it gives insight into the “black box” of schooling. Second, studies of policy effects on student achievement are necessarily complex, involving sophisticated decisions about measurement strategies and analytic approaches that often generate debate. Information on intermediate outcomes – the mechanisms that link a policy initiative to ultimate outcomes -- can increase confidence in estimates of policy effects by providing an empirically based chain of reasoning on how outcomes are generated. And, third, as noted above, because there is generally a lag between the execution of a policy and outcomes, information on these mechanisms can help predict outcomes that are not yet evident. In doing so, it can provide early guideposts to identify likely problem areas that might warrant additional

resources and attention or to caution policy makers not to come to pre-mature conclusions about no effects.

We necessarily cast a broad net for evidence to address these questions. Because state developed accountability systems have been in place longer than NCLB-induced systems, far more evidence is available for effects of state systems than for broadly defined NCLB effects. In this review, we use survey results from teachers and administrators in states that have instituted performance-based accountability policies; national survey data from teachers and administrators; student performance information by subject, as well as qualitative findings and insights gained from small case studies.

1. Do standards and accountability policies lead teachers to focus on different subject areas than they otherwise would?

The basic objective of performance-based accountability policies, including NCLB, is to improve student learning by focusing instructional effort on areas of learning considered important. Defining standards for what students should learn and then conferring rewards and penalties based on how much they have learned is the basic strategy. If the strategy works, one would expect a shift in instruction to subject areas that are identified, tested, and rewarded, resulting in a more focused curriculum. That is the intent. And there is evidence that performance-based accountability leads to an increased focus on raising student achievement in tested subjects. It is notable that the amount of time spent on instruction in mathematics in elementary school increased 40 percent from the 1999-

2000 year to 2003-2004 school years, according to teacher reports in the Schools and Staffing Survey, as state and federal accountability plans were implemented (Hannaway, 2007). Teachers have reported responding to accountability by focusing their efforts more strongly on achievement than they had previously, by working harder, and by seeking to improve their own practices in tested subjects (Bishop & Mane, 1999; Hamilton et al., 2007; Wolf, Borko, McIver, & Elliott, 1999).

At the same time, as noted earlier, not all important areas of learning are tested and incorporated into accountability systems. So there are legitimate concerns about the extent to which standards and accountability policies might distort behavior and *overly* divert attention from worthwhile, but untested, subject areas to tested ones. In other words, the positive changes described in the previous paragraph may be offset by reduced attention to activities and outcomes that are considered important, but that are excluded from the accountability system.

The subjects typically first included in accountability systems are reading/language arts and mathematics. No one would disagree that these are important basic subjects and that a serious level of effort should be given to them. While we do not have representative data on the amount of attention given to these subjects before performance-based accountability policies were enacted, we do have some indication that at least in mathematics the amount of attention given to mathematics instruction in elementary grades varied greatly and was often inadequate.

A study in 1980s of fourth and fifth grade mathematics instruction, prior to performance-based accountability policies, used teacher logs to examine mathematics instruction (Porter, 1989). It found tremendous variation across classrooms in how much instruction occurred: the teacher who taught the most mathematics in this study provided students the equivalent of 23 weeks more instruction in mathematics than the teacher who taught the least (Porter, 1989).² In short, the extent to which students had an opportunity to learn mathematics was largely due to the luck of the draw or to whose class a student might be assigned. The study also found that the preponderance of instructional time in mathematics was devoted to basic skills. Few analysts would find this situation acceptable today. Presumably performance-based accountability in mathematics ensures that teachers provide at least some minimal level of mathematics instruction and also directs them to focus on particular skill sets.

Not surprisingly, the available evidence confirms that, under conditions of a performance-based accountability system, instructional time on tested subjects increases. It also appears that this increase sometimes comes at the expense of non-tested subjects, resulting in a narrowing of the curriculum. We first report the state level findings and then those from the national level. By and large, the patterns are similar.

The question for policy makers, as we review this evidence, is two-fold. First, to what extent is the substitution of tested for non-tested subjects occurring? Is it substantial or is

² A large organizational literature developed in the 1970s and 1980s often referred to as “loosely coupled theory” based on observations about the considerable discretion that resided at the service delivery point in

it small? And, second, is this reallocation worthwhile? The first question is easier to address than the second. Addressing the second question not only involves value issues since individuals differ in the extent to which they value, say, social studies, art, science, etc.; but it also requires some consideration of the possibility that increased capacity in, say, reading or mathematics, may have payoffs not only in reading and mathematics but also in social studies or science. In other words, the amount of social studies a student learns may be a function of both the amount of social studies instruction they receive (holding instructional quality constant) and their reading ability. So if their reading level increases, the amount of social studies they learn may increase even if the amount of time on social studies instruction decreases.

CRESST (Center for Research on Evaluation, Standards and Student Testing) researchers began examining the effects of accountability policies on teachers, mainly in Kentucky and Washington state in the late 1990s. When the research began in 1996, Kentucky was a stand-out leader in state standards and accountability policies. The state tested some subjects in grades 4, 7 and 11 and in other subjects in 5, 8 and 10. Testing only in selected grades is called “milepost testing” and was intended to control and distribute the cost and time burdens of testing.

Milepost testing provided researchers with an opportunity to compare instructional practices in tested grades with non-tested grades and this is what was done by CRESST researchers in Kentucky. By a large margin, Kentucky teachers reported that they

education organizations and the general lack of mechanisms of coordination and control (March and Olsen,

allocated increased time to subjects being tested and decreased the time spent on subjects that were not tested in the grades they taught. Mathematics, for example, was tested in grade 5. Eighty-two percent of 5th grade Kentucky teachers reported increasing instructional time in math; but only 14 percent of 4th grade teachers reported doing the same (Stecher and Barron, 2001). Similarly, science was tested in 4th grade, and 76 percent of 4th grade teachers reported increasing instructional time in science, compared to only 13 percent of 5th grade teachers (Borko, 2005). Estimates of actual hours of instruction by subject also were significantly different between teachers in tested and non-tested). For example, 4th grade teachers, on average, spent nearly 50 percent more time on science instruction than 5th grade teachers (5.2 hours/week vs. 3.5 hours/week) and 5th grade teachers spent more than 30 percent more time on mathematics than 4th grade teachers (6.4 hours/week vs. 4.9 hours/week). So the differences were substantial.

The same researchers also examined shifts in instruction in Washington state that began testing writing, mathematics and reading in grades 4, 7 and 10 in 1998. Researchers conducted two waves of surveys of teachers in grades 4 and 7. Teachers reported shifts in the amount of time they spent on various subjects, and these shifts were consistent with the Kentucky findings described above. In the first survey, about 70 percent of 4th grade teachers reported increasing instructional time on writing and almost 60 percent reported doing the same for mathematics. About half of these teachers also reported decreasing the time spent on social studies, science, and the arts (Hamilton, Stecher, and Klein,

1976;1978; Weick, 1977).

2002; Stecher, Chun, Barron and Ross, 2005). The pattern in the second survey was similar, but not as strong (Stecher and Chun, 2002).

RAND researchers have also been analyzing the effects of accountability policies on instructional practices of mathematics and science teachers in California, Georgia and Pennsylvania, using surveys that were collected in 2004 and 2005 after NCLB had taken hold (Hamilton et al., 2007). Measures of instructional change rely on teachers' retrospective judgments about whether their instructional practices had changed since the prior year. Across the three states, the greatest increases from 2003-2004 to 2004-2005 in time reported by teachers were in core subjects tested according to NCLB accountability requirements – English/ language arts and mathematics (Hamilton et al., 2007). Similar increases were reported in the 2004 data collection. Elementary teachers were more likely than middle school teachers to report reallocating time, a finding that probably reflects the greater flexibility afforded to teachers in self-contained classrooms. This flexibility may also be the reason for the higher likelihood of reductions in time spent on science and social studies than arts or physical education—subjects that are often not under the control of the classroom teacher. It is important to note that the numbers reported in this study may understate actual changes: case studies indicated that in some schools, teachers of non-tested subjects were asked to integrate math or reading into their instruction, resulting in an increase in time spent on math and reading even in the absence of formal scheduling changes.

The findings of Pedulla (2003) and fellow researchers at the NBETPP also suggest test-induced curriculum narrowing. Regardless whether tests were high or low stakes, instruction in tested subjects increased, according to teacher reports, and decreased in non-tested subjects with the strongest effects in the high stakes situations. Still only one-quarter of the teachers working under high stakes testing conditions reported that areas not covered by the tests decreased ‘a great deal’. There is some evidence that the magnitude of the shifts among subjects varies according to the achievement level of the school. Teachers in Colorado, for example, reported that the state’s high-stakes testing system had led to a reduction in time spent on social studies and science, which were not tested, and teachers at low-performing schools were more likely than those at high-performing schools to report these changes (Taylor, Shepard, Kinner, & Rosenthal, 2003).

The instructional shifts that occur in response to performance-based accountability can be prompted by action at any level in the system. Since most evidence on reallocation of instructional time across subjects comes from the reports of teachers,³ many presume that these shifts are a consequence of independent decisions by teachers. But schools and districts also play a role. Using two waves of data from principals in North Carolina, Ladd and Zelli (2002) found that principals reported being significantly more active in 1999 than 1997 in encouraging “greater focus on math and reading in the teaching of other subjects” and in redirecting funds from other subjects to math and/ or reading. Indeed, the increase was more than double for some activities, for example, a little over

20 percent of principals reported redirecting funds in 1997 to over 70 percent in 1999. Similarly, data in Florida show that state accountability policies led to increases in school minimum time policies for different subjects with particularly high bumps in reading, when the accountability system gave it extra weight in its reckoning, and in science when that test was introduced into the accountability system (Hannaway, 2007).

A large multi-year study of the Florida accountability system, that included the full census of schools in the state and that linked administrative data on student performance with survey data on school instructional policies, found that accountability policies contributed to student gains and that a significant part of these gains were attributable to the policy changes (Rouse et al, 2007).

Survey data collected by the Center on Education Policy from the district level administrators in 2005 suggests that time spent on particular subjects in schools may be a consequence of district-developed policies. At least half of the districts in the CEP study require elementary schools to spend a specific amount of time on reading (60%) and on math (50%), and 71% of districts reported reducing time on untested subjects at the elementary school level to provide additional time for reading and mathematics instruction. Urban and high-poverty districts were especially likely to report requiring a certain amount of instructional time on tested subjects, and they also reported spending more time on these subjects than other schools (e.g., urban districts reported requiring an average of 113 minutes per day for reading instruction, whereas the overall average was

³ And, indeed, these are the best reporters since they have the first-hand knowledge of what is happening in

about 90 minutes). (Center on Education Policy, 2006). In addition, district staff often reported policies that required low-performing students to receive extra instruction in reading or math, which usually required the student to miss class time in other subjects.

A detailed analysis of Chicago's accountability policy, instituted in 1996-97, shows similar results. In a variety of ways, including substituting away from untested subjects like science and social studies, teachers responded to the incentives established by the accountability policy. Drawing on the district's individual student level longitudinal database, the Chicago study also examined student performance on the district's high stakes tests and found related results. Gains in math and reading (tested subjects) were two to four times larger than gains in science and social studies (untested subjects), though performance in the untested subjects also increased (Jacob, 2004). Gains were the greatest in low-achieving schools and, interestingly, slightly larger for lower ability students in these schools in the tested subjects and smaller on non-tested subjects, suggesting a resource shift to tested subjects particularly for low achieving students.⁴

It is difficult to explain the increase in performance in untested subjects, but it is possible that there are beneficial spillovers of accountability policies, as we earlier suggested. That is, reading gains may promote gains in other subjects because the written material associated with those subjects may be better understood. It could also be that classrooms facing accountability pressure are overall better organized for learning and this classroom

classrooms.

orientation leads to across the board gains. In short, accountability systems may yield generalized efficiency gains in instruction, even when a limited number of subjects are tested. More detailed evidence and focused analysis are necessary to confirm this observation.

2. *Do standards and accountability policies lead teachers to focus on some types of content and instructional practices over others?*

Reallocation of instructional effort can also occur across skills within subject areas and evidence suggests that the format of the test affects the skills emphasized in instruction. Teachers in Kentucky, for example, where the state test required student written responses, reported increased writing instruction (Stecher et al, 1998). An earlier study in Kentucky also found an increased instructional emphasis on problem solving, as well as writing, as a consequence of the portfolio-based test in that state (Koretz et al, 1996).⁵ This type of reallocation is likely to be driven both by the content included on the tests and the specific ways in which items are designed to measure mastery of that content (for reviews and further discussion of these possible changes see Hamilton, 2003; Koretz & Hamilton, 2006.

⁴ Koretz and Barron (1998) and Dure and Strayer (2001) had similar findings.

⁵ This form of testing, however, is expensive because of the labor involved in grading so its use is limited. A 2003 GAO study estimated that the cost nationally of standardized testing to be \$1.9 billion. If open-ended responses and essays were also included in the format, the cost would be \$5.3 billion. However, new developments in computer-based testing, including the growing feasibility of machine-grading of open-ended responses, could eventually reduce these costs dramatically.

In the NEBPPT national survey, about 7 in 10 teachers reported that state-mandated tests lead teachers to teach in ways that run counter to their ideas of good practice (Pedulla, 2003). Similar responses were obtained from teachers in RAND's three-state study; majorities of teachers disagreed with a statement that the state's accountability system supported their personal approach to teaching (Hamilton et al., 2007). Surprisingly, the Pedulla et al., study indicated that responses for teachers in high stakes environments were not much different from teachers in low stakes environments. Also surprisingly, 81 percent of teachers in high stakes environments reported that state testing influenced them to teach 'critical thinking skills' and about the same percent (83%) reported that testing influenced them to teach basic skills. So while teachers feel that testing diverts their attention in ways that they feel inappropriate, it is sometimes unclear what those ways are.

Evidence from specific state testing programs suggests that the specific kinds of instructional changes teachers report are often a function of the test format (Borko & Elliott, 1999; Lane, Parke, & Stone, in press; Wolf & McIver, 1999). For example, mathematics teachers in Vermont reported increasing their emphasis on problem solving and representations in response to that state's portfolio-based assessment system (Koretz, Stecher, Klein, & McCaffrey, 1994). Although portfolios are rarely used in large-scale testing systems, even shorter, open-ended items have been associated with reported increases in instructional practices that require students to explain their answers and in the use of open-ended tests in the classroom (Taylor et al., 2003; Hamilton et al., 2007). High-stakes writing assessments have been associated with increased classroom time

spent on writing (Koretz, Barron, et al., 1996; Koretz & Hamilton, 2003; Stecher, Barron, Kaganoff, & Goodwin, 1998). Understanding these instructional changes is important for trying to discern the source of gains on statewide tests. Stone and Lane (2003), for example, found that teacher reports of increased use of reform-oriented practices were associated with improved performance on a performance-based assessment in Maryland.

Research examining the more commonly used test formats, particularly multiple-choice items, similarly suggests that teachers respond by emphasizing tested material. Teachers in two districts studied by Shepard and Dougherty (1991) reported increasing time on basic skills, vocabulary, and computation in response to a high-stakes test that was believed to emphasize these skills. Among a national sample of eighth-grade mathematics teachers, Romberg, Zarinia, and Williams (1989) obtained reports of increased coverage of basic skills and computation accompanied by a decreased emphasis on extended projects and other activities not emphasized by most tests. Similar results have been observed in language arts, such as among Arizona teachers who reported de-emphasizing non-tested skills and activities, including writing (Smith et al., 1991). Reports of diminished emphasis on use of essay-based assessments, and of writing instruction designed to emphasize tested activities such as looking for errors in others-written work, have been described in other studies (Shepard & Dougherty, 1991; Darling-Hammond & Wise, 1985). The national study conducted by Pedulla et al. (2003) found that the tendency to adopt instructional materials and practices that mirror the format of state tests is more common in states with high-stakes testing programs than in states with lower-stakes programs. These responses may be of special concern in reading. To the

that extent tests influence teachers to focus on basic reading skills, coupled perhaps with a reduction in social studies, reading comprehension may suffer. Reading comprehension is promoted when students have a broad base of knowledge that they can bring to bear (Zstrow with Janc, 2004).

The effect of high-stakes testing on science has not been studied as extensively as reading and mathematics. As discussed earlier, the presence or absence of a high-stakes science test in a particular grade is associated with the amount of time teachers spend on science instruction. Some observers have concluded that science instruction would benefit from the inclusion of science tests in accountability systems, but even though its inclusion would probably lead to increased emphasis on the subject, there are concerns about the extent to which large-scale science tests measure inquiry and other valued science-related skills and (Pringle and Martin, 2005).

Some analysts suggest that, to the extent reallocation of instructional effort due to the nature of the tests is a problem, the problem is the tests not accountability per se. The notion is that if the right tests were used, instruction would follow, and students would learn the right material. Such reasoning, however, does not take into account the costs associated with alternative tests. In addition, there is evidence that even well-designed tests do not always avoid the problems associated with narrowing and score inflation (Koretz & Barron, 1998). Teachers often continue to engage in narrowed instruction by emphasizing the specific types of problem solving required by the test items (Stecher &

Mitchell, 1995) or by focusing their instruction and evaluation strategies on the rubrics that are used to score the assessments (Mabry, 1999).

3. To what extent do standards and accountability policies lead teachers to focus on skills specific to assessment and testing procedures or to engage in other behaviors that might lead to higher test scores, but have little impact on real learning across students?

Apart from the shifts in instructional practice discussed above, test-based accountability policies may have other effects where the induced behaviors are unlikely to produce real overall learning outcomes across students, even though test scores might rise. Teaching to the test is a prime example. Other examples include targeting particular students or particular subjects that are likely to produce disproportionate gains.

Teaching to the test

Teaching to the test needs to be discussed and defined carefully. While it is often viewed as an adverse consequence of test-based accountability, the basic purpose of test based accountability is, indeed, to direct instructional attention to tested subjects. Although the term “teaching to the test” is sometimes used to encompass all changes in practice that are designed to raise test scores, including increasing time in a subject area, this section deals with the type of teaching to the test in which instructional effort is expended on

skills that increase test scores without increasing the underlying skills and knowledge that the test was designed to measure (Kober, 2002).

Teachers in both nationally representative and state surveys report spending considerable time developing students' test taking skills. (e.g., Quality Counts 2001; Hoffman, Assaf and Paris, 2001). Not surprisingly, this tendency appears to be particularly likely among teachers in states with higher levels of performance accountability pressure (Pedulla et al, 2003). Moreover, teachers' emphasis on test taking is often reinforced by actions of school or district administrators. To illustrate, in one recent study (Hamilton et al., 2007) about 90% of principals in Georgia and Pennsylvania distributed test preparation materials such as practice tests to teachers, and similar percentages distributed released copies of the state tests. Almost all principals said they addressed test preparation in staff meetings and helped teachers identify content that is likely to be on the state test. The pay-off for student learning is unclear. Research suggests, for example, that simply drilling students on items from prior tests, a not uncommon practice among teachers (Popham, 2001), may increase test scores, but is unlikely to lead to real learning. Popham (2001) is careful to distinguish between "curriculum-teaching", where teachers focus on the full body of knowledge and skills underlying a curriculum to be assessed, and "item-teaching" where teachers focus on specific questions likely to show up in a test, but that represent only a small bit of the body of knowledge of interest.⁶ It is the latter teaching to the test that is likely to lead to "test score inflation", "increases in test

⁶ Specific test items are selected to indicate mastery of a broader set of skills and knowledge. To the extent that students learn the answer to a specific item, any inference to the broader knowledge domain is not warranted (Haertel, 1999; Koretz, McCaffrey and Hamilton, 2001).

scores that are not accompanied by commensurate increases in the proficiency levels they are intended to represent” (Koretz, 2003).

However, it is not always easy to determine what kinds of practices fit this description, and there is not necessarily a clear distinction between teaching that improves skills and knowledge and teaching that artificially inflates test scores. Koretz and Hamilton (2006) describe several types of instructional responses to testing. Some of these, such as teaching more effectively, would be considered desirable responses, and one of them—cheating—is clearly undesirable. In the middle are several categories of responses that could be considered ambiguous. For example, coaching on specific types of problems may improve student skills to some degree, but can also lead to score inflation if these skills do not transfer to other problem types. Having students practice filling in multiple-choice answer sheets may improve test validity to the extent that it enables students to demonstrate their skills and knowledge more readily, but beyond a certain amount is likely to be considered undesirable.

A study in Kentucky, an early accountability state, found that students had higher scores on previously used items in the state test than on new items, and further that students in schools with the largest gains had larger discrepancies between new and previously used items. The presumption is that the higher scoring schools used previously used items to coach students (Koretz and Barron, 1998). Other research has also found evidence suggesting that at least some of the gains when tests are re-used may not reflect real learning, at least not fully. For example, Koretz, Linn, Dunbar and Shepard (1991) found

that student performance on a high-stakes exam did not generalize to other exams that were not the focus of the accountability policy. But while “real learning” might not result from “item-teaching”, the incentive for teachers and school leaders to game the system in ways to increase test scores remains. Such behavior is clearly a down-side of test-based accountability systems. It not only wastes precious time with students, but also affects the validity of the information provided by the tests and sends faulty signals to administrators and policy makers about where to focus improvement efforts.

Cheating and other inappropriate behavior

A sure way to increase test score and not increase “real learning” is by outright cheating, whether by students, teachers or administrators. And the stronger the incentive in the accountability system, the greater the incentive to cheat. In Chicago, Jacob and Levitt (2003) found significant increases in outright cheating – changing answers or filling in blank responses -- by teachers or administrators after the introduction of high stakes testing, most prevalently among low-achieving classrooms. The researchers looked for odd patterns of student responses to test items in over 700,000 student tests and estimated that such egregious cheating took place in at least 4-5 percent of Chicago classrooms.⁷

Estimates of the extent of inappropriate teacher behavior also come from surveys of teachers. For example, a substantial minority of teachers surveyed in Maryland and Kentucky reported that teachers rephrased test questions for students during test

⁷ They also estimate that such cheating could only explain a very small part of test score gain since 1996-97 when the Chicago accountability policy took effect.

administration (Koretz, Barron, Mitchell and Stecher, 1996; Koretz, Mitchell, Barron and Keith, 1996). When teachers in two large school districts in Colorado were asked about inappropriate teacher behavior as part of a study of the instructional effects of accountability policies, about 10 percent reported they believed teachers in their schools gave students answers “often” or “frequently” and about 6 percent reported teachers changed student answers (Shepard and Dougherty, 1991). A more recent survey conducted by the National Board on Educational Testing and Public Policy (NBETPP) found fairly low incidence of reported cheating by teachers, even in high stakes environments where, it appeared to be somewhat less likely, perhaps because officials tend to issue strong warnings against cheating and to impose very strong sanctions in these environments.⁸ Only about 1-2 percent of teachers, regardless of the accountability environment, reported that they had heard of teachers or administrators changing scores. The most common report was of teachers giving extra time, ranging from 12 percent in the most high stakes school settings to 19 percent in a lower stakes setting (Pedulla et al., 2003). The results were not available by the performance level of schools so it is unclear how reports vary for higher and lower achieving schools where incentives might well differ. In addition, it is worth noting that obtaining accurate reports about cheating is more complicated than for other behaviors because of the legal issues and strong sanctions associated with cheating.

⁸ Care should be taken in interpreting estimates from this study. While the researchers surveyed 12,000 teachers in high stakes and low stakes accountability environments, the response rate was only 35%. The sample population, however, was similar to the national population of teachers in terms of age, race/ethnicity, type of school (elementary, middle, high school), and experience. See Pedulla, 2003, for details.

Case studies and analysis of school performance in Florida show a different kind of non- or only marginally productive response (Goldhaber and Hannaway, 2004). At the time of the study, schools in Florida received an F grade in the state's accountability system if they failed all three tested subjects: mathematics, reading and writing. Each of these subjects was given equal weight in the accountability, probably in order not to distort instructional behavior. But schools quickly learned that the easiest way to avoid F status was to improve the writing score. Principals reported that they instructed students to write three paragraphs in response to whatever question was given. The first paragraph should begin "first"; the second paragraph should begin "then", and the third paragraph should begin "finally". If students followed this simple rubric, according to principals, students were highly likely to receive a passing grade. Sure enough, none of the 78 school schools that were at risk of receiving a second F received one. While many schools improved in multiple subjects, all the schools passed writing.⁹ This type of rubric-driven instruction clearly led to higher scores, but it is much less clear whether these scores represent a real improvement in students' writing proficiency, especially given the formulaic nature of the instructions given to students.

Students

Another possible inducement of test-based accountability is differential treatment of students either in instructional attention or in testing. Teachers might direct more attention to, say, students near the threshold of some pass point, for example, than

⁹ Interestingly, shortly thereafter, the state announced sweeping changes in the accountability system including decreasing the weight given to writing and increasing the weight given to reading (Goldhaber and Hannaway, 2001).

students either far below the pass point or safely above it. Targeting effort in this way is likely to have the greatest effect on increasing the number of students who make some proficiency standard. It is also in the interest of teachers and administrators to exclude from accountability-related testing students who are likely to do poorly on the test as well as to retain students in grade, thus changing the composition of students taking the test.

A sizable minority of teachers in the California, Pennsylvania and Georgia RAND study reported that they focused more on students who were close to the proficiency mark, suggesting they are diverting attention from other students (Hamilton et al., 2007). Booher-Jennings (2005) provided a rich case study of this practice, which she called “educational triage”. Somewhat in contrast principals in North Carolina reported directing more attention to low-performing students than to high performing students in both 1997 and 1999, though greater instructional effort was reported for both groups of students. The least amount of new attention was given to students who performed on grade level (Ladd, 2001), typical of systems with cut scores. Deere and Strayer (2001) also found evidence of increased attention to low achieving students. At least one recent study shows student achievement gains consistent with an emphasis on students performing near a threshold: Neal and Schanzenbach (2007) analyzed test scores from fifth-graders in the Chicago Public Schools and found that after NCLB was implemented, students performing near the middle of the score distribution showed larger gains than students performing above that level. The gains among the lowest-scoring students were mixed and not as consistent as those of the students scoring near the middle. A similar

pattern was observed after a district-level system, which also emphasized a cut score, was introduced in 1996.

Other studies provide evidence of manipulation of the test-taking population. Jacob (2004) found that while the total proportion of students in Chicago who took the test after the introduction of the accountability policy was about the same as it was before the policy, the percent of students whose test results were excluded for accountability purposes increased, though only modestly.¹⁰ Deere and Strayer (2001) also show increases in special education classification in Texas following the introduction of accountability. Similar findings about the assignment of students to special education have been found by Cullen and Reback (2006) in analysis of Texas data and in Florida data Figlio and Getzler (2002). In both these studies, the students assigned to special education tended to be low performing and minority students.

In addition, in Chicago, Jacob (2004) found evidence of increased grade retention, though Jacob and Lefgren (2002) found a mixed effect on grade retention. And Koretz and Barron (1998) found no evidence on grade retention in their data in Kentucky.

While accountability systems are generally intended to affect mostly students and schools at the low end of the performance distribution, Goldhaber and Hannaway (2004) suggest that there can also be large effects for high performing schools that are not necessarily beneficial. They found 'A' schools in Florida taking actions that were likely not

beneficial for their students' learning, such as delaying all project work and field trips until after spring testing, in order to maintain their A status.

In summary, there is consistent evidence that some of the ways that test-based accountability policies induce teachers and administrators to behave might be considered unethical and make, at best, little contribution to real learning by students or lead to differential effects across students. Clearly efforts should be made to minimize such behavior. Such results, however, should not be used to condemn test-based accountability as a whole. Indeed, the incidence of outright cheating appears relatively low though other inappropriate practices are more common. The questions that need answering are: What is the net effect of test-based accountability policies after taking these negatives into account? And how can accountability systems be designed to minimize the likelihood of negative responses while maximizing the likelihood of practices that are beneficial for student learning?

5. To what extent do standards and accountability policies influence the infrastructure that supports instruction?

Information effects

¹⁰ The treatment of bilingual students is difficult to determine from this analysis because changes in policy for these students were confounded with the introduction of high stakes testing.

Apart from the formal incentive implications of test-based accountability policies, which can lead to benefits as well as unwanted distortions, as discussed above, accountability policies also

produce information that can inform instructional efforts and make teaching more efficient and effective. The idea here is that the information feedback may affect teacher behavior independent of any formally instituted incentive scheme. The presumption is that teachers, like most individuals, prefer to perform their jobs well rather than poorly. They are therefore generally receptive to information that guides their behavior in more productive ways (assuming the costs of acquiring and using the information are not high). This openness to information, of course, is no doubt heightened and focused by accountability policies. The NBETPP survey found that teachers in high stakes testing environments reported that they were more likely to use information from tests to inform their instruction than teachers in low stakes environments, but teachers in low stakes environments also found the information useful (Pedulla et al., 2003).

Other research has also found that performance based accountability policies and the performance information they provide encourage teachers to work harder, be more instructionally focused, and develop their instructional capacities (Wolf, Borko, McIver and Elliott, 1999; Bishop and Mane, 1999). Still information alone may be insufficient to lead teachers to change their instructional behavior in the most productive ways (Firestone et al, 1998). And some types of information might lead to an increase in undesirable responses, such as excessive narrowing of the curriculum. Teachers in high stakes testing environments also reported that they were more likely to use information

from tests to inform their instruction than teachers in low stakes environments (Pedulla, 2003). Support and training are likely to be necessary to ensure that teachers use information in productive ways.

Case study results in Florida, for example, suggest that teachers in low performing schools (F schools) may have been unaware of how poorly they were doing when they were given information comparing their schools with student bodies with similar characteristics, and they reported finding the performance information motivating (Goldhaber and Hannaway, 2004). Teachers also reported that they preferred to be in an F school, where the district allocated additional resources, than in a D school that did not receive such resources, suggesting that their increased motivation was a function of performance information coupled with additional resources to support improvement efforts.

The RAND survey work in California, Pennsylvania and Georgia found that over three-quarters of teachers in each state reported that results of the state mathematics test led them to search for more effective ways to teach mathematics “a moderate amount” or “a great deal” (Hamilton, Berends and Stecher, 2005). When elementary school teachers were asked about whether, as a consequence of the state’s accountability system under NCLB, various aspects of their school had changed for better or for worse, teachers reported that every factor, except one,¹¹ had changed for the better by either a large

¹¹ Far more teachers reported that teacher morale had changed for the worse than for the better: California (56% vs. 7%); Georgia (47% vs. 15%); Pennsylvania (73% vs. 6%).

margin or at about the same rate as reported for changing for the worse.¹² The differences in factors relating to teaching practice were particularly large. For example 10 times more teachers in Georgia reported their own teaching practices changed for the better than changed for the worse (51% vs. 5%) and the patterns in California (33% vs. 8%) and Pennsylvania (38% vs. 15%) while not as large were in the same direction. When asked about ‘teachers’ general focus on student learning’, the pattern of responses was similar with larger percents of teachers reporting ‘change for the better’: Georgia (51% vs. 6%); California (44% vs. 9%); Pennsylvania (39% vs. 15%). The one exception was teacher morale where teachers overwhelming reported that things were worse: Georgia (47% vs. 15%); California (56% vs. 7%); Pennsylvania (73% vs. 6%).

The specific ways that teachers use data from accountability systems, and their receptivity to such data, are likely to be affected by a number of factors including school-level capacity and resources related to data use, opportunities for interaction around data with colleagues, and the specific ways in which the data are provided—particularly their timeliness and accessibility (Coburn, Honig, & Stein, 2005; Marsh, Pane, & Hamilton, 2006; Supovitz & Klein, 2003). Districts that are supportive of data use often take steps to promote school-level use of data. To illustrate, one school district working with the Institute for Learning invested a large amount of resources in the development of a computerized system, accompanied by staff training, to help teachers and other school staff analyze data and use this analysis as input into their school improvement plan.

¹² Responses of middle school teachers were very similar.

Teachers reported that this activity helped them use the data to guide their teaching (Marsh et al., 2005).

One specific way that districts have attempted to promote achievement data use is through the adoption of interim assessment systems that provide teachers and other school staff with frequent, timely information about students' progress toward meeting state standards. A wide variety of interim assessment types has been developed, and some types—particularly those that are integrated into the curriculum and that provide teachers with rich information that can guide teaching and learning while it is happening (sometimes called “formative assessments”)—have been associated with improved student achievement (Black & Wiliam, 1998). In contrast, many of the interim assessments that have been adopted recently are designed to predict state test performance; they tend to be administered less frequently than the formative assessments described earlier, and are less closely tied with day-to-day instructional activities. Many of them have standardized administration conditions and are administered and scored electronically. As noted in a 2005 *Education Week* article, the availability and use of these tools from commercial test publishers has increased rapidly in recent years. There is little research evidence on the effectiveness of these systems, but teachers view them as an important source of information (Hamilton et al., 2008).

Resource allocation

Evidence also suggests that, under some conditions, accountability systems affect the educational infrastructure, in particular financial resource allocation and staffing patterns,

in ways likely to support instructional practice. For example, results show that during the early days of standards-based reform, districts serving higher poverty populations increased spending on instructional support service proportionately more than districts serving more affluent populations, possibly in an effort to increase student performance (Hannaway and Nakib, 2002). But results also show that early accountability states only increased investments in instruction if they also increased spending overall (Hannaway, McKay and Nakib, 2002). State policies appear to have large effects. In Florida, for example, major reform efforts were accompanied by increased expenditures for instructional staff support and increases for professional development were greater in high poverty and low performing schools (Hannaway, Nakib, Fu and Stanislawski, 2005).

CONCLUSIONS

The empirical findings to date suggest the incentives associated with accountability systems have powerful behavioral implications for classrooms and schools. Some of these effects are negative; other appear to be beneficial. Cheating and item teaching are clearly negative responses to test-based accountability and appear to be stimulated by test-based accountability. In addition, time to non-tested subjects, such as art, social studies and foreign languages appears to be sacrificed in favor of tested subjects. The extent to which this substitution is negative or beneficial is debatable and partly value based.

Other classroom level effects appear to be beneficial. Standards and aligned tests are shaping instruction and, at the least, they insure that the subjects considered by many to be the most critical subjects, particularly mathematics and reading, get considerable classroom attention. There is also some evidence that performance based accountability policies and the performance information they provide encourage teachers to work harder, be more instructionally focused, and develop their instructional capacities.

The power of accountability systems means that their details are important and demand careful attention and caution from policy makers. The strong influence of standards and especially tests on instructional practice places a significant burden on those standards and tests to be high quality. In addition, the targeting of particular students or cutoff scores can also make a significant difference in how teachers and administrators respond.

Research findings should also be treated cautiously. Accountability systems change over time as unintended consequences and other effects become evident to policy makers. In short, we should be careful about inferring too much from the cross sectional results of accountability systems. A dynamic model is needed to examine feedback and subsequent behavior both of policy makers and the subjects of accountability pressure - local administrators, teachers and students.

References

- Bell, B. & Cowie, B. "Formative Assessment and Science Education". Dordrecht, Netherlands: Kluwer Academic Publishers. In Ruiz-Primo, M. A. & Furtak, E. M. *Informal Formative Assessment of Students' Understanding of Scientific Inquiry*. Center for the Study of Evaluation (CSE Report # 639). Los Angeles: University of California, 2004.
- Bishop, John and Ferran Mane. *The New York State Reform Strategy: The Incentive Effects of Minimum Competency Exams*. Philadelphia, PA: National Center on Education in Inner Cities, 1999.
- Black, P. and Wiliam, D. "Inside the Back Box: Raising Standards through Classroom Assessment." *Phi Delta Kappan* 80.2 (1998): 139-148.
<http://www.pdkintl.org/kappan/kbla9810.htm>.
- Bloom, B. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Education Researcher* 13.6 (1984): 4-16.
- Cavanagh, Sean. "Students Double-Dosing on Reading and Math." *Education Week*. 14 June 2006.
- Center on Education Policy. "From Capital to the Classroom." March 2006. <http://www.cep-dc.org/nclb/Year4/CEP-NCLB-Report-4.pdf>
- Center On Education Policy Symposium. "Forum on the Costs and Legal Issues Surrounding the Implementation of the No Child Left Behind Act." 14 July 2005. <http://www.cep-dc.org/pubs/nclby3/forumJuly2005.htm>
- Crooks, T. J. "The Impact of Classroom Evaluation Practices on Students." *Review of Educational Research* 58.4 (1998), 438-481.
- Cullen, Julie Berry and Randall Reback. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." Working paper, University of Michigan, 2002.
- Deere, Donald and Wayne Strayer. "Putting Schools to the Test: School Accountability, Incentives, and Behavior" Texas A&M (2001)
- Dillon, Sam. "Test Shows Drop in Science Achievement for 12th Graders." *The New York Times*. 25 May 2006.
- "Doubling up on Literacy Classes." Editorial. *The New York Times*. 2 April 2006.
- Elmore, Richard and Robert Rothman, eds. *Testing, Teaching, and Learning: A Guide for States and School Districts*. National Research Council. 1999.

Figlio, David and Lawrence Getzler. "Accountability, Ability and Disability: Gaming the System." NBER Working Paper W9307. Cambridge, MA: National Bureau of Economic Research. 2002.

Firestone, William, et al. eds. *The Ambiguity of Teaching to the Test: Standards, Assessment, and Educational Reform*. LEA. 2004.

Firestone, William, et al. "Performance-Based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland." *Educational Evaluation and Policy Analysis*. 20.2 (1998): 95-113.

Fontana, D., and Fernandes, M. "Improvements in Mathematics Performance as a Consequence of Self-Assessment in Portuguese Primary School Pupils." *British Journal of Educational Psychology* 64.3 (1994): 407-417.

Frederiksen, J.R., and White, B.J. "Reflective Assessment of Students' Research within an Inquiry-Based Middle School Science Curriculum." American Educational Research Association Conference, Chicago, IL. 1992

Fuchs, L., Fuchs, et al. "Effects of Task-Focused Goals on Low-Achieving Students with and Without Learning Disabilities." *American Educational Research Journal* 34 (1997): 513-43.

Goldhaber, Dan and Jane Hannaway. "Accountability with a Kicker: Observations on Florida Vouchers". *Phi Delta Kappan*, 85 (8): 598-605, 2004.

Grissmer, David and Ann Flanagan. "Tracking the Improvement in State Achievement Using NAEP Data." RAND Corporation, unpublished paper. 2002.

Haertel "Performance Assessment and Education Reform." *Phi Delta Kappan* 80 (1999): 626-666.

Hamilton, Laura. "Assessment as Policy Tool." *Review of Research in Education*, Vol 27 (2003): 25-68.

Hamilton, Laura, Mark Berends and Brian Stecher. "Teachers' Responses to Standards-Based Accountability." Santa Monica, CA: RAND., 2005.
http://www.rand.org/pubs/working_papers/2005/RAND_WR259.pdf.

Hamilton, L.S., Stecher, B.M., Marsh, J., McCombs, J.S., Robyn, A., Russell, J., Naftel, S., & Barney, H. (2007). *Implementing standards-based accountability under No Child Left Behind: Responses of superintendents, principals, and teachers in three states*. Santa Monica, CA: RAND.

Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. In B. Fuller, M.K. Henne, & E. Hannum (Eds.), *Strong state, weak schools: The benefits and dilemmas of centralized accountability (Research in the Sociology of Education, Vol. 16, pp.31-66)*. St. Louis, MO: Emerald Group Publishing.

Hannaway, Jane “Higher Order Thinking, Job Design, and Incentives: An Analysis and Proposal,” *American Education Research Journal*. 29.1 (1992): 3-21.

Hannaway, Jane. “Unbounding Rationality: Politics and Policy in a Data Rich System”. Misticfer Lecture, the University Council of Education Administration, November 17, 2007.

Hannaway, Jane and Maggie Stanislawski. “Responding to Reform: Florida School Expenditures in the 1990s”, *Journal of Education Finance*, under review,.

Hannaway, Jane, Shannon McKay and Yasser Nakib, "Reform and Resource Allocation: National Trends and State Policies". In Wm Fowler (ed), *Developments in School Finance, 1999-2000* (edited by William Fowler). U.S. Department of Education, 2002.

Hoffman, James, Lori Czop Assaf, and Scott Paris. “High-Stakes Testing in Reading: Today in Texas, Tomorrow?” *Reading Teacher*, 54.5 (2001): 482-492.

Holmstrom, Bengt and Paul Milgrom. “Multi-Task Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,” *Journal of Law, Economics and Organization*, 1992.

Jacob, Brian. “Remedial Education and Student Achievement: A Regression-Discontinuity Analysis.” *Review of Economics and Statistics* LXXXVI.1 (2004): 226-244.

Jacob, Brian and Steven Levitt. “Rotten Apples: An Investigation of the Prevalence and Predictors of Teachers Cheating.” NBER Working Paper 9413. Cambridge, MA: National Bureau of Economic Research. 2003.

Jerald, Craig. ““Teach to the Test’? Just Say No.” Center for Comprehensive School Reform and Improvement Issue Brief, 2006.
<http://www.centerforcsri.org/files/CenterIssueBriefJuly06.pdf>

Jones, M. Gail, Brett Jones, Belinda Hardin, Lisa Chapman, Tracie Yarbrough and Marcia Davis. “The Impact of High-Stakes Testing on Teachers and Students in North Carolina.” *Phi Delta Kappan*. 81.3 (1999): 199.

Kober, Nancy (2002) “What Tests Can and Cannot Tell Us.” *Testtalk for Leaders*, 2 (2002).
<http://www.cep-dc.org/testing/testtalkoctober2002.pdf>

Koretz, Daniel, Sheila Barron, Karen Mitchell, and Brian Stecher. “The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS).”

- RAND Education Publication MR-792-PCT/FF. Santa Monica, CA: Rand Corporation. 1996.
- Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.531-578). Westport, CT: American Council on Education/Praeger.
- Koretz, Daniel, Karen Mitchell, Sheila Barron, and Sarah Keith. "The Perceived Effects of the Maryland School Performance Assessment Program." CSE Tech. Rep. 409. Los Angeles: Center for Research on Evaluation, Standards and Student Testing. 1996.
- Koretz, Daniel "Attempting to Discern the Effects of the NCLB Accountability Provisions on Learning." Paper presented at the annual meeting of the American Educational Research Association, Chicago. 2003.
- Koretz, Daniel and Sheila Barron. *The Validity of gains on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: The Rand Corporation. 1998.
- Koretz, Daniel, Robert Linn, Stephen Dunbar and Lorrie Shepard. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests." Paper presented at the annual meeting of the American Educational Research Association, Chicago. 1991.
<http://www.colorado.edu/education/faculty/lorrieshepard/PDF/High%20Stakes%20Testing.pdf>
- Ladd, Helen, ed. *Holding Schools Accountable: Performance-Based Reform in Education*. Washington, DC: The Brookings Institution Press. 1996.
- Lazear, Edward. "Speeding, Tax Fraud, and Teaching to the Test." *NBER*. 2004.
<http://www.nber.org/papers/w10932>.
- Loveless, Tom. "Test-Based Accountability: The Promise and the Perils." *Brookings Papers on Education Policy: 2005*. Washington, DC: Brookings Institution Press: 2005. 13-14.
http://muse.jhu.edu/journals/brookings_papers_on_education_policy/v2005/2005.1loveless.pdf
- Mathis, William. "The Cost of Implementing the Federal No Child Left Behind Act: Different Assumptions, Different Answers." *Peabody Journal of Education*, 80.2: 90-119.
- Mehrens, William and John Kaminski. "Methods for Improving Standardized Test Scores: Fruitful, Fruitless or Fraudulent?" *Educational Measurement: Issues and Practice*, 8.1 (1989): 14-22.

Meisels, Samuel, Sally Atkins-Burnett, Yange Xue, Julie Nicholson, Donna DiPrima Bickel, and Seung-Hee Son. "Creating a System of Accountability: The Impact of Instructional Assessment on Elementary Children's Achievement Test Scores." Education Policy Analysis Archives. 11.9 (2003).

<http://epaa.asu.edu/epaa/v11n9/>

Milgrom, Paul and John Roberts. "Adaptive and Sophisticated Learning in Normal Form Games." *Games and Economic Behavior*, 3.1 (1991): 82-100.

National Research Council. *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington D.C.: National Academies Press, 2001.

National Research Council. *Classroom Assessment and the National Science Education Standards*. Washington, D.C.: National Academy Press, 2001.

Newmann, Fred, Anthony Bryk, and Jenny Nagaoka (2001) "Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?" Chicago: Consortium on Chicago School Research, 2001.

<http://www.consortium-chicago.org/publications/pdfs/p0a02.pdf>

Organization for Economic Co-operation and Development. *Improving Learning in Secondary Classrooms*. Centre for Educational Research and Innovation, 2005.

Pedulla, Joseph, Lisa Abrams, George Madaus, Michael Russell, Miguel Ramos, and Jing Miao. *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers*. Boston: National Board on Education Testing and Public Policy. 2003.

<http://www.bc.edu/research/nbetpp/statements/nbr2.pdf>

Peyser, James and Robert Costrell. "Exploring the Costs of Accountability." *Education Next*, 4.2 (2004): 23-29.

Popham, W. James. "Teaching to the test." *Educational Leadership*, 58.6 (2001): 16-20.

Porter, Andrew. "A Curriculum Out of Balance: Elementary School Mathematics," *Educational Researcher* 18 (1989): 9-15.

Pringle, Rose and Sarah Martin. "The Tests Are Coming: The Impact of Upcoming High-Stakes Testing on the Teaching of Science in Elementary Classrooms." *Research in Science Education*, 35.2 (2005): 1-15.

Quality Counts 2001: A Better Balance. Education Week, 2001.

Rodriguez, M. "The Role of Classroom Assessment in Student Performance on TIMSS." *Applied Measurement in Education* 17.1 (2004): 1-24.

- Romberg, Thomas, E. Anne Zarinnia and Steven Williams. *The Influence of Mandated Testing on Mathematics Instruction: Grade 8 Teachers' Perceptions*. Madison, WI: National Center for Research in Mathematical Sciences Education. 1989.
- Rose, Lowell & Alec Gallup. "The 36th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools." *Phi Delta Kappan*, 86.1 (2004): 41-56. <http://www.pdkintl.org/kappan/k0409pol.pdf>
- Rosenbusch, Marcia Harmon. "The No Child Left Behind Act and Teaching and Learning Languages in U.S. Schools." *The Modern Language Journal*, 89 (2004): 250-261.
- Rouse, Cecilia, Jane Hannaway, Dan Goldhaber and David Figlio. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure", CALDER Working Paper #13, November, 2007.
- Sadler, D.R. Formative Assessment and the Design of Instructional Systems. *Instructional Science* 18.2 (1989): 119-144.
- Shepard, Lorrie. "Inflated Test Score Gains: Is the Problem Old Norms or Teaching to the Test?" *Educational Measurement: Issues and Practice*, 9 (1990): 15-22.
- Southern Regional Education Board. "Raising Achievement and Improving Graduation Rates: How Nine High Schools That Work Sites Are Doing It." June 2005. http://www.sreb.org/programs/hstw/publications/briefs/05V14_ResearchBrief_raising_graduation_rates.pdf
- Stecher, Brian and Sheila Barron. "Quadrennial Milepost Accountability Testing in Kentucky." CSE Technical Report 505. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 1999. <http://www.cse.ucla.edu/Reports/TECH505.PDF>
- Stecher, Brian and Sheila Barron. "Unintended Consequences of Test-Based Accountability When Testing in 'Milepost' Grades." *Educational Assessment*. 7.4 (2001): 259-281.
- Stecher, Brian, et al. "The Effects of the Washington State Education Reform on Schools and Classrooms: Initial Findings." RAND Education, 2005. http://www.rand.org/pubs/working_papers/2005/RAND_DB309.pdf.
- Stecher, Brian, et al. "Introduction to First-Year Findings from the Implementing Standards-Based Accountability (ISBA) Project." *RAND Education*. 2005. http://www.rand.org/pubs/working_papers/2005/RAND_WR255.pdf.

Stecher, Brian, et al. "Measuring Reform-Oriented Instructional Practices in Mathematics and Science." *RAND Education*. 2002.
<http://www.rand.org/pubs/drafts/2005/DRU2787.pdf>.

Stiggins, R. & Conklin, N.F. *In Teacher's Hands: Investigating the Practices of Classroom Assessment*. Albany, NY: SUNY Press, 1992

Stiggins, R. *High Quality Classroom Assessment: What Does it Really Mean?* NCME Instructional Topics in Educational Measurement Series, Module 12.
<http://www.ncme.org/pubs/items.ace>

Taylor, Grace, Lorrie Shepard, Freya Kinner, and Justin Rosenthal. *A Survey of Teachers' Perspectives on High-Stakes Testing in Colorado: What Gets Taught, What Gets Lost*. Boulder: University of Colorado, School of Education, Education and the Public Interest Center, 2001.
<http://www.education.colorado.edu/epic/coloradostudiesandreports.htm>

United Kingdom. Office for Standards in Education. "Evaluating Mathematics Provision for 14-19 year olds." May 2006.

von Zastrow, Claus. "Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools." (Washington, DC: Council for Basic Education, 2004).
<http://static.ncss.org/files/legislative/AcademicAtrophy.pdf>

Wolf, Shelby, Hilda Borko, Monette McIver, and Rebekah Elliott. "'No Excuses': School Reform Efforts in Exemplary Schools of Kentucky." CSE Tech. Rep. 514. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 1999.
<http://www.cse.ucla.edu/Reports/TECH514.pdf>