Commentary: Accountability Policy and Scholarly Research

Frederick M. Hess, American Enterprise Institute

Since 2001, considerations of school reform have been dominated by performance-based accountability. No Child Left Behind (NCLB) has changed the way policymakers and educators talk about education, look at educational performance, and think about educational challenges. Nonetheless, NCLB and the state accountability systems it has spawned have been subjected to little careful scrutiny. This article discusses four recent research contributions and considers how they might inform policymaking on accountability. While scholarly scrutiny will not necessarily settle debates, it can help yield more constructive and informed decisions. In particular, research can clarify the actual consequences of policy decisions; highlight and refine approaches that may be more reliable, stable, and effective than those in use; flag the unanticipated or overlooked effects of design decisions; and ensure that both policymakers and the public are aware of the costs and benefits of accountability.

Keywords: accountability, testing, No Child Left Behind, policy

Since 2001, considerations of school reform have been dominated by performance-based accountability. No Child Left Behind (NCLB) has, for better or worse, changed the way policymakers and educators talk about education, look at educational performance, and think about educational challenges. Nonetheless, NCLB and the state accountability systems it has spawned—in all their awkward compromises, complexities, and jury-rigged determinations—have been subjected to limited careful scrutiny. Far more common have been cheery promises of educational improvement or shrill complaints about the dangers of testing. There has been a preference for stylized story-telling and selected factoids and a lack of interest in level-headed assessments of what accountability systems are doing, how they are operating, or how we might sensibly refine them.

There have been, however, some useful efforts to assess the components of the NCLB testing, assessment, and accountability system. A number of organizations have done admirable work in tracking various dimensions of NCLB implementation. While each of these efforts has its flaws and is shaped in subtle or more obvious ways by the perspective of its organizational sponsor, collectively, they have provided useful overviews of how the testing, accountability, and remedy provisions of NCLB have been implemented (see, for instance, Center on Education Policy, 2005; Education Commission of the States, 2002; Education Trust, 2004). These invaluable studies have been largely descriptive and narrative, presenting information on state activity, student achievement, and implementation schedules. This work has focused on presenting data rather than assessing competing implementation strategies or examining the effects of particular policy decisions.

A few scholars have examined the state role in the implementation of NCLB, addressing the challenges of assessment, measurement, and accountability (Goertz, 2005; Wanker & Christie, 2005). Other research has considered the law’s sanctions and remedy provisions in “developed” accountability states like Michigan and Florida, though such analyses invariably involve examining the design and modification of accountability systems (Hannaway & Bischoff, 2004; Plank & Dunbar, 2004). Peterson and Hess (2005) have documented the immense disparity in the rigor of state accountability systems, and the perverse reality that NCLB’s inadequate yearly progress (AYP) requirements make school performance look worse in states with more demanding accountability systems. Hoxby (2005) and Walberg (2005) have provided suggestions on how to make NCLB testing and assessment more effective and reliable. Popham (2005) has also provided a useful, critical look at the design and execution of the NCLB system. While these analyses provide an important first step, much more research is essential to fully understand how states are complying with the NCLB requirements or the significance of the design decisions they are making as they do so.

The four papers collected in this volume provide an important contribution to this growing body of research. They constitute some of the most thoughtful efforts to date to describe how NCLB-mandated accountability is being put into effect, what kinds of operational decisions this entails, how those decisions are affecting results, and how these systems are affecting educational outcomes more generally.

The allure of performance-based accountability is easy to understand. Standards represent a public commitment that schools will teach all children a discrete body of knowledge and skills to a specified level of mastery. Setting meaningful performance standards, however, makes it inevitable that
some students, teachers, and schools will fail to meet those standards. This poses a daunting political challenge in public schools, where “low performers” have powerful incentives to voice their displeasure and challenge the legitimacy of the accountability system.

These critiques attract substantial sympathy and typically have some merit. After all, performance-based accountability rests on key assumptions, many of them drawn from experiences in other sectors. Because accountability tools have not traditionally been used in schooling, refined for an educational context, or applied directly to K–12 management, a number of practical and hypothetical questions arise in the design and implementation of high-stakes accountability. This lack of experience creates concerns both about how well these systems will translate and about the challenges of sustaining them in the educational sector.

Accountability systems, such as those mandated by NCLB, require public officials to make a number of politically sensitive decisions when it comes to testing, designing sanctions, and determining adequate school performance. States must design or adopt a test and designate a passing score for each grade level and subject. They must determine whether adequate performance will be based on the level of achievement or the rate of student growth. They must determine the minimum number of students who will be tested for a group score to be considered reliable and valid. They must determine how variables other than measured student achievement will be factored into evaluations of school performance. Depending on the accountability in question, a slew of similar such decisions is typically required.

Proponents have difficulty standing firm on the details of any particular accountability system because the essential components of determining what is tested, where passing scores should be set, how scores for schools should be calculated, and what sanctions or rewards should be attached to specified outcomes are inherently arbitrary. They are public policy decisions that ultimately rest on judgment, compromise, and political determinations. The closer one gets to crafting and enforcing the specific decisions required for an accountability system, the less defensible any one particular program element may appear. Though accountability is a useful artifact, determining what students need to know, how well they need to know it, and whether schools are effectively teaching it, is an ambiguous and value-laden exercise. Because public schooling requires public officials to make these judgments and impose them statewide, these difficult questions inevitably become political ones (Hess, 2002).

Critics seize upon the arbitrary nature of many of these decisions, raising concerns and arguing for modifications that they believe will increase test validity and reduce inequities or pernicious effects produced by the misuse of assessments. Of course, the reality is that most such changes also happen to benefit the influential educational constituencies (e.g., the group of districts, educators, or students) that the critic represents. Empirical analysis cannot wave away these concerns or alleviate the need to come up with imperfect solutions for each of these challenges.

However, scholarly research of the kind conducted by the authors in this special issue can do four important things. First, it can clarify the actual consequences of policy decisions, putting to rest exaggerated claims by both proponents and critics of various decisions. Second, it can suggest and refine approaches that may be more reliable, stable, and effective than those in place. Third, it can flag the effects of mundane decisions, helping officials recognize when minor design decisions have had large or unanticipated effects. Finally, it can ensure that both policymakers and the public are aware of the costs and benefits of accountability—and how both of these are distributed.

### Reviewing the Relevant Provisions of NCLB

Even at this late date, several years after NCLB was enacted, there is still significant confusion about the law and what it requires—both out in the field and even in Washington, DC. In fact, the law is relatively prescriptive on some elements of implementation but quite ambiguous on many others, making the kind of research presented here useful not just as an academic exercise but also for crafting state implementation statutes and revising federal guidelines.

The NCLB accountability system requires states to develop content standards for what their students are expected to know, create assessments that reflect these standards, and annually test to measure performance on the standards in the “core subjects” of reading and mathematics (and eventually science). The expectation is that states will reach “universal proficiency” in reading and mathematics by 2013–14, with all students meeting state content standards. What “universal proficiency” means is left to each state to determine. With regard to assessments, the law gives little concrete guidance. It requires that state plans “demonstrate that the State has adopted challenging academic content standards and challenging student academic achievement standards” but makes clear that these standards need not be approved by the U.S. Department of Education.

With regard to testing student mastery of these standards, however, the law becomes much more prescriptive. States must administer tests in reading and mathematics every year in grades 3–8 and at least once in high school. States are also required to design and administer an annual science assessment by 2007–08 and to test students with it at least once in grades 3–5, 6–9, and 10–12. The reading, mathematics, and science assessments must “cover the breadth and depth of state content standards,” be “reliable and of high technical quality,” and report student results in terms of the required state achievement standards (e.g., basic, proficient, and advanced). Again, what all this means in practice is a determination left largely up to the states.

AYP is the metric by which all schools and school districts are evaluated under NCLB, and the law’s fundamental requirement is that all schools and districts “make AYP.” The concept of “making AYP” can be likened to schools and districts jumping over a bar. The bar is the percentage of children that must score “proficient” on mathematics and reading assessments. States had a large degree of flexibility in determining where to first set the bar, but over time, it must be continually raised, so that by 2013–14, it is set at 100%. This means, for instance, that a state could deem schools to be making AYP in 2005 if 40% of students were “proficient” in reading, but expectations would have to be ramped up to 100% by 2013–14.

One can see that a state with a tough definition of proficiency and a high bar for AYP would identify many more schools as needing improvement than a state with a weak definition of proficiency and a low bar.
for AYP. In fact, by 2005, some states had virtually no schools identified as needing improvement while other states identified close to 80% of theirs. A corollary is that “universal proficiency” will have very different meanings in different states.

The law is quite specific, however, about how states are to analyze the data that determine which schools and districts made AYP. NCLB requires that AYP determinations be statistically reliable and valid; be based primarily on state academic assessments; and include another indicator, like graduation rates for high schools or attendance rates for elementary and middle schools. Because AYP is based primarily on academic achievement, gains on this second indicator are not enough to make schools clear the AYP bar. However, poor performance in graduation rates or attendance can result in a school failing to make AYP, even if its achievement levels are satisfactory.

In calculating AYP, NCLB requires that states hold schools accountable for the achievement of a variety of subgroups within each school, including all racial/ethnic groups present in the school (white, African American, Latino, Native American, etc.); low-income students; students with disabilities; and students with limited English proficiency. For a school to make AYP, each subgroup must clear the bar in reading and in mathematics.

NCLB anticipates that at some schools these subgroups will be quite small. In order to protect student confidentiality, as well as ensure statistical reliability, states are allowed to set the minimum size, also known as “n size,” under which subgroup scores will not count. As Porter, Linn, and Trimble demonstrate in this volume, the “n sizes” range from 5 to 100, with most clustering around 30 or 40. Differences in these provisions obviously have an enormous impact on determining whether schools will make AYP.

The Four Papers

The four studies here complement one another, combining to provide a portrait of how decisions shape the impact of accountability systems and how accountability systems, in turn, influence student achievement. The Chester and Porter et al. studies look squarely at the decisions involved in crafting state-level accountability systems and at the consequences of those decisions. The Desimone, Smith, Hayes, and Frisvold study and the Carnoy study explore how pre-NCLB accountability systems have affected student achievement and graduation rates. Together, the studies make clear both how much we have to learn and how much even preliminary analyses may help to discipline and clarify our thinking.

Chester’s discussion of Ohio’s accountability system provides a primer on how other states might shape their systems to improve reliability and stability. Rather than substituting NCLB for the existing state accountability system, Ohio opted to incorporate the NCLB measures alongside its own system in a manner designed to ensure that schools were being judged fairly on their overall performance, their treatment of vulnerable populations, and their rate of improvement. Chester’s depiction of Ohio’s sophisticated system, drawing upon multiple performance measures to place school performance into one of five classifications, shows how much seemingly technical decisions may matter. Chester explains how the move from a system that relied entirely upon a handful of “performance indicators” to the new system relieved the pressure on schools to concentrate on students who were “on the bubble” of proficiency and strengthened public support for the accountability system. This detailed explanation of the system and presentation of how frequently the various performance metrics influenced the ranking of schools and districts can help inform thinking in those states considering similar measures.

Porter et al. begin by providing descriptive data on state policies of improvement timelines, minimum subgroup size, and the use of confidence intervals to protect against measurement error. The data are themselves immensely useful, are not widely known, and help make clear which states have been more or less aggressive in pursuing school-level accountability. Porter et al. then conduct an analysis of Kentucky’s achievement data, which reveals the impact these three technical questions actually have in determining whether schools make or miss AYP. The results demonstrate that seemingly innocuous rules governing minimum “n size” and confidence interval sizes turn out to make an enormous difference in the number of schools identified as in need of improvement under NCLB. For instance, in 2003, just 10% of Kentucky schools failed to make AYP, under rules which included a confidence interval and only counted scores for subgroups of at least 60 students. Without changing the passing score or any other substantive elements of the test, dropping the confidence interval adjustment and counting scores for subgroups of at least 30 students would have caused 55% of schools to fail to make AYP. Differences like that matter a lot, both for policy and for the maintenance of political support.

Not surprisingly, given how much these refinements may improve the apparent performance of a state’s school system, many states have engaged in gamesmanship that respects the spirit of NCLB, while exploiting whatever leeway the law offers. Porter et al. note that about half of all states have chosen to “backload” the increases in performance expectations that will move all states to 100% proficiency by 2013–14, while not a single state has opted to “frontload” the increases. This means that most states have required schools to make only modest achievement gains in the early years under NCLB, while expecting heroic gains in the later years. Because states were also largely free to decide where to set their initial proficiency bar, critics have noted that NCLB can reward states that set a lower bar for themselves and thereby inflate their apparent performance. In shining a bright light on these perverse incentives and making clear how much “technical issues” actually matter, Porter et al. illustrate the need for school reformers to keep the pressure on state officials. Their results also suggest the need for federal officials to address the manner in which the perverse incentives produced by NCLB’s flexibility may undermine the position of those states eager to comply with the spirit of the law.

In their article, Desimone et al. take an ambitious look at whether states tend to enact accountability systems in a coherent or piecemeal manner and how accountability systems affect student learning. The reality is, as Desimone et al. acknowledge, that we do not yet have data to examine these questions as systematically or rigorously as we would wish. The authors do find that states which embrace some components of a highly developed accountability system—such as coherent
standards and efforts to align curricula to standards—are more likely to adopt other components. Particularly helpful is their determination to move beyond simple dichotomies and to examine accountability systems as comprehensive “systems,” composed of a variety of discrete elements, and to then determine how various constellations of program elements affect outcomes. This approach provides a useful contrast with analyses that imply that accountability systems are either present or not, while ignoring the reality that accountability systems can come in all kinds of shapes and sizes.

The results presented by Desimone et al. are intriguing, raising the likelihood that the effects of accountability will depend on the kinds of content, students, and schools in question. More important, however, is the promise of the data set and analytic infrastructure that the authors have started to construct. Combining data from a number of existing sources and creating a sophisticated, coherent set of variables that can be merged with emerging streams of achievement data, Desimone et al. have helped to develop a model that will provide increasingly helpful information on the effects of accountability.

Finally, in an effort that has certain similarities to the Desimone et al. analysis, Carnoy uses a thoughtful definition to classify state accountability systems and then examines the impact of state accountability on high school graduation rates. Contrary to the more fervent claims that proponents of high-stakes graduation testing make, Carnoy finds minimal effects. Proponents of high-stakes testing have long asserted that such tests will raise standards and encourage students to take school more seriously, thereby helping to boost graduation rates; critics have argued that they cause schools to hold low-achieving students back, discourage them, and thus reduce graduation rates for vulnerable populations. Examining data from all fifty states, Carnoy notes that accountability appears to have a slight, negative correlation with the rate at which lower-income minorities graduate high school, but he notes that it is not clear what this finding means. Meanwhile, the graduation rates of both white and minority students have increased in the past seven years, a period during which accountability systems have been widely adopted.

More fundamentally, Carnoy makes clear the limits of empirical analysis for making policy regarding accountability. He notes that any negative results might be just a short-term effect on graduation rates, and that the longer-term effect might prove to be positive, as his previous research in Texas found. The nature of accountability is that the short-term costs for some students or schools may be highly visible, while the longer-term benefits may take years to show up and be difficult to measure. In short, even if research findings point in a particular direction, we cannot be sure that the results tell us everything—or even that future results will look the same.

Public Opinion and the Workings of NCLB

One reason that these questions of program design and the impact of accountability matter so much is the importance of public support in determining the fate of accountability systems. Accountability is a simple and pretty rugged idea, but it rests on the delicate, subjective determinations of policymakers and public officials. These decisions are inevitably judgment calls and can never be defended on purely “objective” or “scientific” criteria. Consequently, the reliability and durability of accountability systems finally depend on the public’s willingness to accept them as reasonable, sensible, and workable. How particular decisions play out will matter greatly in shaping public opinion.

This tension in public thought is illustrated when Americans are asked about the principles underlying NCLB and the law itself. For instance, the 2005 Educational Testing Service (ETS) poll found that 55% of adults believed “all students, teachers, and schools should be held to the same standard of performance,” while 34% of adults disagreed (Hart & Winston, 2005). However, other surveys suggest discomfort with uniform standards in practice, particularly when it comes to students with special needs or limited English proficiency. The Public Education Network’s (2004) poll found that 58% of voters disapprove of holding special education students to the same standards as other children, and 46% disapprove of holding limited English proficiency children to the same expectations. The 2004 PDK/Gallup poll reported that 61% of respondents did not believe that special education students should be held to the same standard as other students (Rose & Gallup, 2004).

Analysis as presented in this issue is so useful in part because public opinion remains so mixed in regard to NCLB. Quite simply, the public has not yet made up its mind on whether NCLB is workable or likely to achieve its goals. The ETS reported in 2005 that 45% of adults had favorable opinions toward NCLB and 38% had unfavorable opinions. Among those who had strong views, opinion was evenly split, with 19% of all adults holding a strongly favorable opinion and 21% a strongly unfavorable one. The ETS also reported what it termed a “worrisome ... disconnect” between the views of adults, on one hand, and teachers, on the other. While adult views of the law were mixed, 50% of high school teachers held a strongly unfavorable opinion, and just 2% held a strongly favorable one (Hart & Winston, 2005).

What Empirical Work Can and Cannot Do

The four analyses presented here are a significant contribution to our understanding of NCLB and performance-based accountability. It is necessary to be clear, however, on what this kind of work can and cannot reasonably be expected to accomplish. Empirical scholarship cannot ultimately determine which systems of accountability are necessarily “best” or even “prove” to skeptics that performance-based accountability is desirable. Those who believe that educators are much like other people and school systems suffer from the same organizational inertia that other kinds of large organizations do will find accountability an essential device for forcing hard decisions, driving performance, and fostering a culture of competence (Hess, 2003). Those skeptical of behavioralist assumptions or who deem teachers and schools fundamentally different from most other employees and organizations will continue to regard accountability as a threat to the integrity of teaching and learning (Kohn, 1999).

The thoughtful analyses presented here, and others like them, will probably not convince many attentive readers that performance-based accountability is a good idea or a bad idea. Such judgments are informed by life experience,
perspectives on human nature, research and insight from many different sectors, and similar factors that stretch beyond scholarly inquiries into the workings of educational accountability. This is true as much for officials in Washington as for teachers and principals in the nation’s schools.

That said, this work provides invaluable fodder for debating how educational accountability is working in practice, understanding the impact of design decisions, and shaping the broader public’s understanding of the costs and benefits of performance-based accountability. Rather than hope this research will settle disputes, offer a tranquil respite from partisan strife, or reveal the “one-best” design for accountability systems, we should understand that its value resides in its ability to inform the debates and illuminate ambiguities. For instance, Chester’s analysis does not determine the “best” way to synthesize measures into a comprehensive accountability system. However, it provides state policymakers with useful guidance on more and less sensible ways to configure such systems. Similarly, Porter et al. cannot tell policymakers whether particular guidelines would constitute good policy, but the research can open their eyes to the consequences of the choices they need to make.

Debates over accountability are sorely lacking in empirical measures of what is actually transpiring. State and federal officials are struggling to make sensible decisions about minimum subgroup size, the number of metrics used to judge schools, where to set performance cut-offs, how much confidence to put in test results, and many similar questions. They are tackling these issues in an atmosphere fraught with technical challenges, political pressures, and concerns about equity. Under such conditions, what scholars can do best is explain the choices and try to anticipate their consequences. Such analysis will be essential if we are to forge educational accountability systems that are effective, stable, and backstopped by broad popular support.

References


