The Nelson A. Rockefeller Institute of Government

The policy research arm of the State University of New York

411 State Street
Albany, NY 12203-1003
(518) 443-5522

www.rockinst.org

EDUCATION POLICY BRIEF

April 2008

Structural Problems in Educational Accountability

In October 2007, the Rockefeller Institute of Government brought together a group of some 40 state and federal education officials, testing experts, educational researchers, and policy advocates for a symposium on intergovernmental approaches for strengthening K-12 standards and test-based accountability systems, with support from the Spencer Foundation and the Joyce Foundation. Later this year, Institute co-director Richard Nathan and Allison Armour-Garb, director of education studies, will publish a final report on this project. The following article contains excerpts from a background paper that was circulated to participants in advance of the symposium.

By Allison Armour-Garb

Six years into the implementation of the No Child Left Behind Act (NCLB), many state agencies lack adequate access to, or budgets to pay for, the expertise they need to implement and monitor sound accountability systems. Policymakers, educators, and testing companies face incentives to cut corners, lower standards, and game the system, and the public lacks a clear idea of the effectiveness of the various components of the education system because curriculum standards and measures of performance vary widely from state to state.

These are not problems that can be worked out by tweaking isolated policies, such as when a standard or cut score is set too high or too low. Rather, they are "structural"[1] features of the educational accountability sector that probably require changes in institutions and incentives.

Some argue that the federal government should address these problems by establishing national standards and tests. But state governments and other stakeholders and experts are leery of federal control. In the meantime, a number of states are collaborating to develop and use common standards and improve testing systems, by participating in consortia such as the American Diploma Project Network, the State Collaborative on Assessment and Student Standards, and the New England Common Assessment Program.

This article focuses on the structural problems in educational accountability that are motivating not only these collaborations but an increasing number of proposals for new intergovernmental mechanisms.

1 Guidelines for designing and evaluating accountability systems — "a new endeavor"[2]

According to the basic theory of educational accountability systems, in return for the public dollars they receive, educational entities at the school and district levels must show that they are achieving certain outcomes. Standards- and test-based accountability systems define those outcomes primarily in terms of student performance on tests that are tied to curriculum standards established by the state. By attaching consequences to performance, accountability systems are intended to drive improvements in educational processes.

In view of education policymakers' heavy reliance on test-based accountability, the effectiveness or "validity" of such systems deserves close scrutiny. Some skeptics question the theory of action behind standards-based reform, arguing that low-performing schools lack the capacity to improve. Others are inclined to support the theory but want to know which variants are most effective.

While experts have a long history of examining the validity of tests (defined as the extent to which the assessment measures the knowledge or skills that it is intended to measure), efforts to establish the validity of accountability systems are relatively recent.[3] The validity of an accountability system has been defined as the degree to which: "[1] The components of the system are aligned to the purposes, and are working in harmony to help the system accomplish those purposes; and [2] The system is accomplishing what was intended" (and has not had unintended negative consequences).[4]

The validity of the tests themselves is a precondition for the validity of any test-based accountability system.[5] Test validity is evaluated in accordance with the *Standards for Educational and Psychological Testing* (often simply called the *Test Standards*),[6] which are jointly developed and periodically revised by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

NCLB has spurred the development of a new literature on the validity of educational accountability systems. Organizations such as the National Center for Research on Evaluation, Standards, & Student Testing, the Consortium for Policy Research in Education, the Center for Assessment, and the Council of Chief State School Officers have collaborated to publish standards and evaluation frameworks for state accountability systems (see box on page 3, "Evaluating Accountability Systems: An Emerging Consensus"), but these have not yet achieved the canonical status of the *Test Standards*.

2 Shortage of Expertise

Creating and implementing a high-quality accountability system that conforms to the *Test Standards* and the newer accountability standards is difficult, labor-intensive, and expensive. These standards are highly technical and can be properly followed only by experts trained in measurement theory and statistics,[7] known as psychometricians, only a few of whom enter the field each year.[8]

Evaluating Accountability Systems: An Emerging Consensus

Despite the lack of a single authoritative set of standards for accountability systems, leading experts agree that policymakers must evaluate the following:

- Goals and theory of action of the accountability system—How is the system supposed to work?
- Indicators used to make accountability decisions—What is being measured?
- Accuracy and consistency of the decision rules used to classify schools— Were the "right" schools identified?
- Consequences—Were the intended benefits realized? Were the costs (in terms of unintended negative consequences) minimized?
- Interventions—What rewards, sanctions, and supports were implemented? What was effective?

Sources: Eva L. Baker et al., "Standards for Educational Accountability Systems," *CRESST Policy Brief* 5 (Winter 2002), comment to standard 21; Scott Marion, "Evaluating the Validity of State Accountability Systems: Examples of Evaluation Studies," (presentation, American Educational Research Association conference, San Diego, CA, April 15, 2004), slide 3; Ellen Forte Fast and Steve Hebbler, *Validity in State Accountability Systems,* Implementing the State Accountability System Requirements Under the No Child Left Behind Act of 2001 (Washington, D.C.: Council of Chief State School Officers, February 2004), 78.

The surge in demand for testing under NCLB has led to a critical shortage of psychometricians. (See box on page 4, "What Does NCLB Require?") Because states and districts across the country are all mandated to comply with the law's testing mandates, the need for psychometricians is widespread and decentralized. Unfortunately, federal policymakers imposed these new requirements without a full appreciation of the technical challenges they would pose, both for the professionals who would write the tests and the laypeople charged with implementing them.[9] Of course, many of the government officials who write and implement education laws are education experts, broadly speaking, but relatively few trained psychometricians work in the education bureaucracy at any level. Not only does this help to explain the weaknesses in NCLB's test-based accountability provisions, it is a problem that should be overcome before enactment of future such policies.

The shortage of expertise has troubling consequences. Psychometricians who work for private companies typically receive much higher salaries than those who work in the public sector. Those who work as consultants or in academia may have the freedom to choose their own projects, collaborators, and work schedules.[10] Education agencies — which typically offer neither competitive salaries nor flexible working conditions — have therefore been hit hardest by the shortage and have experienced high turnover in these positions.

Due to this difference in hiring power, many state education departments lack staff able to design a technically sound testing program. Most states purchase their tests from commercial test publishers, and many states rely on outside consultants for advice. Yet some states lack sufficient technical know-

how to supervise outside testing contractors effectively.[11] When states work with independent testing advisors for just a few days per year, officials may not know the right questions to ask, or — not grasping the significance of consultants' recommendations — may fail to implement them.[12]

This results in what economists call "asymmetric information." The officials who write education laws and buy tests, due to a shortage of experts on their payrolls, may not be aware that their state accountability systems are running afoul of professional standards. Even if the testing companies are producing technically defensible tests, the government's uses of the test results or the conditions of administration may violate professional norms. The affected students and the public are even less likely to understand the ways in which educational accountability systems may be flawed.

What Does NCLB Require?

NCLB requires states to implement statewide accountability systems based on challenging state standards and annual testing in reading and mathematics for all students in grades 3-8. Each state must establish its own definition of "proficiency" and must report assessment results broken out by poverty, race, ethnicity, disability, and limited English proficiency to ensure that no group is left behind. Each state must set annual statewide progress objectives ensuring that all groups of students reach the "proficient" level on state tests by the 2013-14 school year.

Individual schools must meet state "adequate yearly progress" (AYP) targets toward this goal for both their student populations as a whole and for each demographic subgroup. School districts and schools that fail to make AYP will, over time, be subject to improvement, corrective action, and restructuring measures aimed at getting them back on course to meet state standards.

A sample of 4th and 8th graders in each state must also participate in the National Assessment of Educational Progress testing program in reading and math every other year to provide a point of comparison for state test results.

Source: U.S. Department of Education, "NCLB Overview," http://www.ed.gov/nclb/overview/intro/execsumm.html.

3 Perverse Incentives and a Focus on Compliance

Before the implementation of NCLB, politicians, education officials, teachers, and test publishers stood to gain in varying degrees from increases in student test scores, whether or not those scores represented real gains for children. But NCLB has increased the stakes — it added both carrots and sticks — and thereby increased the need for effective oversight of educational accountability systems. On the incentive side, the law ties states' federal Title I funds to compliance and has made billions of dollars available to test publishers. It also increases the visibility of high-scoring districts and those that post proficiency gains for all students and disadvantaged subpopulations. The sticks are the negative consequences (programmatic and political) that flow from failing to show "adequate yearly progress" in the percentage of children who score proficient on the required tests each year.

NCLB has posed new financial, practical, and technical challenges for both testing companies and states, and with those challenges come strong incentives. The testing market is dominated by highly competitive for-profit corporations whose primary motive may be to profit by satisfying client demands. NCLB has greatly increased the amount of criterion-referenced testing that states must conduct, and it is difficult and expensive for test publishers to create and score so many new tests within NCLB's tight timelines. Moreover, each state sets its own curriculum standards and is supposed to tailor or "align" its tests to those specific curriculum standards. The fact that testing companies must custom-design aligned tests for each state has multiplied the challenges that states and testing companies face in implementing NCLB.[13] The norm-referenced tests that were more common in the past did not require customizing or annual "refreshing" and were therefore much cheaper to produce.

These new demands are "squeezing testing company profit margins" [14] and have created pressure to cut corners — to design instruments that test lower-order thinking and are generic (i.e., weakly aligned) so they can be marketed to multiple states with minimal customizing.[15] These companies may seek, to a greater or lesser extent, to follow professional standards, but the threat of formal professional censure (which is rare) is far less salient than the desire to secure the largest possible client contracts and get the testing done and scored on time.

Similarly, education officials at the federal, state, district, and school levels — whether they are politicians, psychometricians, or anyone else connected with a particular jurisdiction's accountability systems — have strong incentives to demonstrate increases in the percent of students scoring at the proficient level on their watch, while avoiding increases in taxes or budgets. As a result, they may be unlikely to push testing companies for improvements that could result in longer timelines, higher price-tags, or more challenging test questions.[16] In addition to such inaction, educators can take a more active approach — "gaming" the system by teaching to the test or focusing on students whose scores are just below the cut score, allowing students extra time, or even cheating outright.[17]

One of the easiest ways for states to avoid or postpone sanctions under NCLB is to set a low bar for proficiency. This approach is possible because states, districts, and schools face sanctions for failing to make "adequate yearly progress" (AYP) towards their state's proficiency standards under NCLB, while states are free to set proficiency standards high, low, or in between. States that set high standards risk having the most schools labeled "failing" under NCLB. Indeed, many states set very low proficiency standards (compared to the NAEP) to begin with, and others lowered their standards after they recognized the incentives to do so. This type of behavior is sometimes called a "race to the bottom."[18] Thus, NCLB may have in some cases exerted downward pressures on state proficiency standards.

What of legal incentives? How does the law deal with those who do not adhere to professional standards in their design and implementation of educational accountability systems in the public schools? The answer is that, for the most part, it does not. The *Test Standards* and newer accountability standards are not laws or regulations; they are professional guidelines. Some aspects of the *Test Standards* are paraphrased in the NCLB legislation and regulations, and states' adherence to these provisions is overseen through a peer review process. (See the text box on page 6, "NCLB Peer Review.")

NCLB Peer Review

Under NCLB, states are required to prepare a report documenting aspects of their accountability systems and to submit it for peer review evaluation. The Education Department selects the peer review teams, which typically consist of a psychometrician, an educator who is an expert in working with special populations, and another testing professional with experience in large-scale testing. Employees of testing companies are excluded from the teams. This system provides for expert, in-depth review of some aspects of state accountability systems. The teams' decisions are not "all or nothing," and states have an opportunity to request technical assistance.

Some analysts have noted that peer review has tended to focus on inputs and processes (e.g., how much are states spending on NCLB tests?) rather than on the consequences of state systems (e.g., what is the impact on student learning? Are there unintended negative consequences, such as narrowing in the range of cognitive skills or subjects areas taught, or a "dumbing down" of tests?).

Sources: Brian Gong, Testimony to the Commission on *No Child Left Behind,* www.aspeninstitute.org/atf/cf/%7BDEB6F227-659B-4EC8-8F84-8DF23CA704F5%7D/Brian%20Gong%20Testimony.pdf; Wayne J. Camara and Suzanne Lane, "A Historical Perspective and Current Views on the *Standards for Educational and Psychological Testing*," *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006), 20; Phoebe Winter, telephone conversation with author, September 19, 2007; Susan L. Davis and Chad W. Buckendahl, "Evaluating NCLB's Peer Review Process: A Comparison of State Compliance Decisions" (paper presented at 2007 annual meeting of National Council on Measurement in Education (NCME)); Susan L. Davis, e-mail message to author, Sept. 8, 2007.

In general, however, the testing profession has a tradition of self-regulation. Psychometricians who do not follow the *Test Standards* may be subject to professional censure, but that sanction is imposed rarely and is a weak counter to the financial motives that test publishers face. Moreover, when tests are misused by the laypeople charged with implementing them, professional censure is not an available check. In sum, there is no adequate mechanism for enforcing the *Test Standards*, much less the newer standards for accountability systems.

As a result, educational accountability policies rarely are subjected to serious debate or scrutiny. Officials, educators, and test publishers face incentives to cut corners to save money, and to "game the system" to show short-run increases in the percent of students scoring proficient on their watch — i.e., during their term of office, during the term of their contract, or during their employee review period. The prime incentive facing the federal government, the states, and the testing companies is to do what it takes to make NCLB compliance work, at least for the time being, in order to keep the Title I money flowing.

4 Lack of Transparency

The variation in state standards and tests makes it difficult to interpret the test score data that states generate. Because each state defines its own curriculum standards, uses its own tests, and sets its own bar for proficiency, it is impossible to directly compare student learning and proficiency rates across states. And because some states have changed their proficiency standards from one year to the next, it is also difficult to compare proficiency rates over time. Moreover, some states have systems for grading or ranking schools and districts that are at odds with the federal system, which adds to the public confusion.

The only reliable yardstick for comparing children's educational achievement across states and over time is the NAEP. Numerous studies and reports have sought to compare state standards with the NAEP, and even to map state test scores onto the NAEP scale,[19] but such studies are not yet routine, nor have they reached the majority of ordinary citizens.[20]

5 Inefficiencies Resulting From Diversity of Standards

Variation in state standards and tests also poses challenges for teachers, schools of education, test developers, and textbook companies when they are deciding what material to cover, and for students when they move from one state to another.

As noted above, because state standards vary widely, publishers of tests and textbooks theoretically ought to custom-develop materials for each state. They save on development costs and increase profits, however, by developing generic materials that cover the common elements of multiple states' standards. In the case of tests, publishers may do some minimal customizing. It is questionable, however, whether such weakly aligned tests are effective at measuring how well students have learned what they were taught.[21]

Similarly, because of the variation in standards across states, teachers are not necessarily prepared to teach to the standards in the state in which they teach. They may attend a teacher education program in one state and move to another after graduation.

6 More Research Needed on Validity of State Accountability Systems and Tests

To some degree, the lack of scrutiny of state accountability policies is due to a scarcity of hard data on their direct and indirect impacts. Experts believe that the gap between the goals of educational accountability systems and the actual strength of the research base supporting policy reforms has led to problems.[22]

Obviously, when problematic test results are used to make high-stakes decisions about individual students, such as whether to promote them or grant them a high school diploma, the effects are harmful. But poorly designed or implemented accountability systems are also harmful when they lead education officials to base decisions about curriculum, resource allocation, personnel, or sanctions on flawed information.

With the increased stakes attached to educational accountability systems, there is a correspondingly greater need for routine evaluations of those systems, using criteria like those listed in the box on page 3, "Evaluating Accountability Systems: An Emerging Consensus."

Unless states and districts evaluate their systems, they will have no evidence to defend their accountability decisions or the imposition of sanctions, and they risk losing credibility with stakeholders.[23]

While a few states have years of usable data on their state systems and have been open to researchers, independent validity studies are far from routine in most states. Ongoing evaluations and longitudinal studies would provide better information on which policymakers could base their decision making.[24]

7 Conclusion

Although these structural problems came to light as states, districts, and schools implemented NCLB, simply repealing the law would not solve the problems. Any system of educational accountability — whether at the federal, state, district, or school level — will have to contend with these issues.

The Rockefeller Institute and others are actively exploring how states can work together most effectively to address these problems. In October 2007, the Institute convened a group of some 40 state and federal education officials, testing experts, educational researchers, and policy advocates for a symposium on intergovernmental approaches for strengthening K-12 standards and test-based accountability systems. (An article about the symposium by Lynn Olson and an edited transcript are available here.)

Symposium participants considered possible models for an intergovernmental collaborative, ranging from existing state-led consortia to a new federal agency modeled on the Consumer Product Safety Commission. They talked about how different models balance federal, state, and local interests and discussed the role of the private sector (i.e., employers, testing organizations, foundations). What functions might such an intergovernmental entity perform? Should it develop common standards for certain subjects and grade levels? Common tests? Should it primarily support state efforts by providing technical assistance and research, or is stricter oversight of accountability systems needed?

Many states are already demanding greater commonality, particularly with respect to what students should know and be able to do when they graduate from high school and enter higher education or the workplace. In light of the 2007 extension of NCLB without amendments, there is a window of opportunity now for interested parties to consider options, develop coalitions, and work on new intergovernmental approaches to strengthen educational accountability in the post-NCLB era.

Endnotes

[1] Thomas Toch, Margins of Error: The Education Testing Industry in the No Child Left Behind Era (Washington, D.C.: Education Sector, 2006), 6.

[2] Brian Gong, "Documenting Validity of Accountability Systems: A Progress Report" (presentation, Council of Chief State School Officers National Conference on Large-Scale Assessment, San Antonio, TX, June 20, 2005), slide 2.

[3] Gong, "Validity of Accountabiity Systems," slide 2.

[4]Scott Marion et al., *Making Valid and Reliable Decisions in Determining Adequate Yearly Progress*, Implementing the State Accountability System Requirements Under the No Child Left Behind Act of 2001 (Washington, D.C.: Council of Chief State School Officers, December 2002), 38.

[5]Scott Marion, "Evaluating the Validity of State Accountability Systems: Examples of Evaluation Studies," (presentation, American Educational Research Association conference, San Diego, CA, April 15, 2004), slide 4.

[6] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, D.C.: Authors, 1999).

[7] Robert L. Linn, "Following the *Standards*: Is it Time for Another Revision?" *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006), 54-55; Koretz, telephone conversation with author, June 20, 2007.

[8] Toch, Margins of Error, 9.

[9] William G. Harris, "The Challenges of Meeting the *Standards*: A Perspective from the Test Publishing Community," *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006): 43.

[10] Rather than work full-time for a single testing company or government agency, many testing experts opt to offer their services as consultants to multiple agencies. For example, they serve on states' "technical advisory groups" or "technical advisory committees"; work with state consortia such as the American Diploma Project Network, the State Collaborative on Assessment and Student Standards, and the New England Common Assessment Program; or provide technical assistance and review through the U.S. Department of Education's NCLB peer review process, its Assessment and Accountability Comprehensive Center, and its LEP (Limited English Proficiency) Partnership.

The multistate perspective that such arrangements give many testing experts may be seen as a fortuitous consequence of the national shortage of psychometricians. If every state could afford to hire and retain its own staff of experts, those experts would likely develop a more bureaucratic and parochial mindset.

[11] Toch, *Margins of Error*, 9, 20.

[12] Lauress Wise, interview with author, July 20, 2007.

[13] Harris, Challenges of Meeting the Standards, 43; Toch, Margins of Error, 9.

[14] Toch, Margins of Error, 12.

[15] Ibid, 14-16.

[16]NCLB may also inhibit states from experimenting with innovations in accountability design. Note that although federal pressures and sharing of best practices across state borders has led to the dissemination of growth models, the original value-added model emerged in Tennessee during a period of independent state action.

[17] Evidence of test misuse is abundant. See Wayne J. Camara and Suzanne Lane, "A Historical Perspective and Current Views on the *Standards for Educational and Psychological Testing*," *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006), 38.

[18] According to one new report, there has not actually been a "race to the bottom," with the majority of states dramatically lowering standards under pressure from NCLB, but rather a "walk to the middle," as some states with high standards dropped their expectations toward the middle of the pack. John Cronin et al., *The Proficiency Illusion* (Washington, D.C.: Thomas B. Fordham Institute and Northwest Evaluation Assoc., October 2007) 30.

[19] E.g., National Center for Education Statistics, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales (NCES 2007-482)* (Washington, D.C.: U.S. Department of Education, June 2007).

[20] With the possible exception of those who watched the June 11, 2007, episode of "The Colbert Report" on Comedy Central, which explained the methodology using Mississippi's 4th grade reading test as an example.

[21] Toch, *Margins of Error*, 14-16.

[22] Eva L. Baker and Robert L. Linn, "Validity Issues for Accountability Systems," in *Redesigning Accountability Systems for Education*, eds. Susan H. Fuhrman and Richard F. Elmore (New York: Teachers College, 2004), 47.

[23] Ellen Forte Fast and Steve Hebbler, *Validity in State Accountability Systems,* Implementing the State Accountability System Requirements Under the No Child Left Behind Act of 2001 (Washington, D.C.: Council of Chief State School Officers, February 2004), *77*.

[24] Camara and Lane, "Views on the *Standards*," 39; Toch, *Margins of Error*, 14 (quoting Scott Marion); Daniel Koretz, telephone conversation with author, June 20, 2007.