

Stability and Change in Title I Testing Policy



LORRAINE M. MCDONNELL

This article examines the history of Title I's student testing requirements, focusing on the two purposes they have served as a policy tool and measurement instrument. It argues that these purposes have been defined by a stable core of testing requirements whose specific targeting and technical characteristics have evolved in response to changes in Title I's institutional and interest group environment. In concluding, it considers the form and purpose that the testing requirements are likely to take in the future of the Elementary and Secondary Education Act.

Keywords: Elementary and Secondary Education Act (ESEA), student testing, federal K–12 education policy

As the cornerstone of federal education policy, Title I of the Elementary and Secondary Education Act of 1965 (ESEA) has a multifaceted history spanning five decades and extending from Congress to the classroom (for early histories, see Bailey and Mosher 1968; McLaughlin 1975; Graham 1984; Murphy 1991; and Kirst and Jung 1991; for later analyses, see McDonnell 2005; Manna 2006; McGuinn 2006; DeBray 2006; Cohen and Moffitt 2009; and Rhodes 2012). This article revisits some of that history through the lens of how Title I's student testing provisions have evolved from 1965 to the present. It focuses on the elements that have remained basically the same, those that have changed, and the factors that explain this combination of stability and change.

Program requirements specifying which students should be tested, how they should be assessed, and how the results should be used have served two distinct but related purposes.

First, like most major federal policies, Title I includes provisions designed to act as politi-

cal and policy instruments. They aim to make targets' behavior consistent with federal policy goals, and typically operate through a combination of incentives, regulations, and bargaining. Although the choice and use of these instruments depend on goals articulated in congressional and executive branch policies, their ultimate effectiveness in a federal system is shaped by a broad array of interests and institutions extending from Washington to local classrooms. Requiring states and school districts to test students and to evaluate local programs as a condition of Title I funding has been, together with financial reporting requirements, the federal government's main tool for promoting its goal of improving the educational opportunities afforded low-income students. As a policy tool, testing can fulfill several functions. It may reduce information asymmetries between the federal government and the street-level where Title I services are delivered by providing information (albeit imperfect) about the program's effectiveness.

Lorraine M. McDonnell is professor of political science at the University of California, Santa Barbara.

Direct correspondence to: Lorraine M. McDonnell, mcdonnell@polsci.ucsb.edu, University of California, Santa Barbara, Department of Political Science 9420, Santa Barbara, CA 93106.

Such information can mobilize program constituents to take action in support or opposition to the status quo, and test results often become evidence used in debates about future program directions.

At the same time, these testing requirements also function as a technical instrument for measuring student outcomes. In this second role, the student assessments are judged by established psychometric standards of reliability, validity, and fairness. Is student achievement—however defined—being measured consistently across students and over time? Is a test actually measuring what it purports to measure, and are the conclusions and inferences drawn from the test results appropriate? Is a test systematically underestimating the skills and knowledge of a particular group (National Research Council 1999a, 71–72)?

Throughout the history of ESEA, these two purposes have rarely operated independently of each other and, in some cases, have posed direct trade-offs among them. For example, the policy uses of testing have often yielded reactive effects because educators have organized their teaching to improve their students' scores on required assessments (Stone 2012, 198). In these instances, policy purposes may preempt educators' professional judgments in using testing to guide their teaching. Similarly, when policymakers have used test results for purposes for which an assessment has not been validated, they compromise it as a measurement instrument. The ways in which these two purposes have interacted with each other over time have depended on how policymakers, educators, and interest groups have applied them. The result has been that though the overall purposes of the testing requirements have remained constant, some aspects of how they have been operationalized have changed but other elements have remained stable. The next three sections examine this pattern of enduring policy purposes combined with changing strategies for pursuing them.

STABILITY IN TITLE I TESTING POLICY

As Title I's testing provisions have been operationalized as a policy tool and measurement instrument, three aspects have remained stable throughout its history: a focus on student

assessment as central to an evaluation and accountability strategy, testing as a tool to leverage state and local practice, and a constituency with testing and evaluation as part of its advocacy strategy even though its membership has changed over time.

Evaluation and Accountability Through Student Assessment

At one level, the path to testing as a central component of Title I began with a seemingly small event. Senator Robert Kennedy (D-NY) made clear to the Johnson administration's ESEA architects that he viewed the educational problems of poor children as partly due to the inability and unwillingness of local school districts to address their needs. Kennedy indicated that his support for the ESEA legislation depended on the addition of a reporting requirement that would hold educators responsible for educational achievement as the major criterion in judging ESEA's effectiveness. Part of that accountability strategy was to make information available to parents about how their children were doing. As a result of Kennedy's ultimatum, language was included in the original ESEA legislation requiring that "effective procedures, including provisions for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally deprived children" (ESEA, Title I, sec. 205 (5)).

As Milbrey McLaughlin notes in her analysis of the early history of evaluation in Title I, Kennedy viewed evaluation as a political accountability strategy (1975, vii). The widespread provision of information as a basis for holding public agencies accountable and as a resource that those affected by a policy can use in making decisions and in taking action was not a common policy instrument at the time of ESEA's passage. However, it has now become a customary element in environmental, consumer finance, and health-care policies as well as in education (Fung, Graham, and Weil 2007; McDonnell 2004). But the effectiveness of these hortatory or transparency policies ultimately depends on the quality of the information provided. Informational quality is where ESEA's

evaluation requirement as a policy tool intersects with testing as a measurement instrument. The data available to the U.S. Office of Education (USOE) in the early years of Title I were often spotty and anecdotal. The quality and methodological approaches of local district reports varied considerably and in many cases, districts ignored state requests for student achievement results or provided incomplete data (Borman and D'Agostino 2001, 26).

This failure to provide reliable and valid data stemmed partly from the lack of technical expertise in local school districts. However, as histories of ESEA's early years have documented, much of the low quality and variation were due to ESEA's vague legislative language and the political circumstances surrounding its passage (McLaughlin 1975; Graham 1984; Kirst and Jung 1991; Murphy 1991). Competing expectations on the part of policymakers and interest groups about whether ESEA would be the first step in general federal aid or a program specifically targeted on underachieving, low-income students had been successfully sidestepped in the vague statutory language that aided its quick congressional passage. Inexperience on the part of USOE and state education agencies in managing a large grant program and their political vulnerability unprotected by clear legislative intent meant that Title I's initial implementation was characterized by "compromise and ambiguity" (Kirst and Jung 1991, 46). Even after four successive reauthorizations of ESEA between 1965 and 1980 specified more precisely that Title I should be used to assist educationally disadvantaged students and USOE increased its monitoring, the focus was on fiscal accountability, not programmatic substance. This focus was not surprising, given that Title I remained "more a funding mechanism than a specific program or policy for helping at-risk students" (Vinovskis 1999, 189).

Assessment as a Tool for Leveraging State and Local Practice

Despite the shortcomings of ESEA's testing and evaluation provisions, they have continued on the path begun with the legislation's initial enactment. The primary reason is that these regulations have been the strongest tool

available to the federal government for leveraging state and local practice. Given that its limited constitutional authority and funding status in education make it the proverbial junior partner in the federal system, the federal government has few tools available to advance its social policy goals among state and local targets. As Helen Ingram argued more than thirty-five years ago, rather than buying compliance by offering grants-in-aid such as ESEA, the federal government really only purchases the opportunity to bargain with the states (1977). It exerts limited authority over state and local uses of ESEA funds through fiscal targeting requirements and subsequent audits. However, in the federal government's attempts to influence educational programs delivered to Title I students, the testing provisions have been among its strongest bargaining chips. Even if it had known what instructional strategies are most effective in educating low income students, a provision in the original ESEA statute—still applicable today—prevents the federal government from prescribing that level of programmatic detail: "Nothing contained in this Act should be construed to authorize any department, agency, officer, or employee of the United States to exercise any direction, supervision, or control over the curriculum, program of instruction, administration, or personnel of any educational institution or school system" (ESEA, Title VI, sec. 604).

Consequently, the federal government has had to rely primarily on requiring ex post reporting of program results rather than prospectively mandating or even guiding the organization of classroom teaching. The effect has been that the Title I testing requirements have created an enormous system of state and local testing, and they also launched the development of educational evaluation as a research specialization (U.S. Congress 1992; National Research Council 1999b; Shepard 2008). As we see in subsequent sections, Title I's modest effects, documented in several national evaluations, eventually led to changes in ESEA. Nevertheless, the basic principle of requiring that students be regularly tested in reading and mathematics on a standardized assessment has persisted. A major reason has been the extent to which the early statutory language, in-

cluded to secure ESEA's enactment, precluded alternative forms of federal leverage.

Testing and Evaluation as Part of an Advocacy Strategy

At the same time, a coalition of interest groups has reinforced the continuation of the testing requirements as an accountability mechanism to monitor whether local Title I programs are serving their intended beneficiaries effectively. The membership of this coalition has shifted over time. During ESEA's early years, organizations such as the National Welfare Rights Organization, the NAACP, and the Lawyers Committee for Civil Rights Under Law advocated on behalf of low-income students by pushing for increased federal monitoring to ensure that Title I funds were spent to meet their educational needs (Kirst and Jung 1991). Over time, the groups and the focus of their advocacy changed. Organizations pressing for greater accountability expanded to include the Business Roundtable and the National Alliance of Business as well as newer organizations with an equity agenda such as the Education Trust (DeBray-Pelot 2007; Rhodes 2012). What has been constant throughout ESEA's history is that these groups have supported enforcement of Title I's categorical requirements in counterpoise to traditional education interest groups that, though now supporting Title I's social policy goals, have advocated for greater state and local flexibility in program administration.

The explanation for how testing became a core part of Title I and why it has endured are linked. The original impetus is an example of a seemingly small, contingent event, but the reasons for its continuation are due to historical forces extending beyond just ESEA. Robert Kennedy's amendment established testing and evaluation requirements as a central element of Title I. As states' and local districts' past history of segregation and disregard for poor chil-

dren had demonstrated, his distrust of their likely use of federal funds and his efforts to institutionalize a partial remedy were well placed. Local districts' expenditures of ESEA funds during its early years further buttressed the view of those advocating on behalf of low-income students that federal monitoring was necessary. Because the primary concern during ESEA's early days was ensuring that funds were spent on appropriate program targets and less on which goods and services were purchased, fiscal monitoring took precedence over testing to measure student outcomes (Jennings 2001, 14). Nevertheless, during ESEA's first fifteen years, eight evaluations of Title I were conducted based on student test data to produce national estimates of the program's effectiveness (Borman and D'Agostino 2001).¹

Although its original inclusion in Title I can be explained by Kennedy's amendment, the stability of testing as a key policy tool for federal leverage is best explained by the institutional factors that define education policy in the U.S. federal system—namely, the federal government's limited formal authority and an ingrained political culture legitimating state and local autonomy along with the variation it produces. The centrality of testing requirements in Title I is a case of strong path dependency in which institutional characteristics fundamental to the nature of the American state have made the costs of diversion from that path politically and administratively prohibitive. However, as policy ideas, testing technology, and political dynamics have shifted, the configuration and direction of that path have also been altered.

CHANGES IN TITLE I TESTING POLICY

Although Title I has maintained its essential policy goals and basic categorical structure for fifty years, its program rules have been significantly altered. The testing and evaluation requirements have been central to this transfor-

1. Carl Kaestle and Marshall Smith, writing in 1982, argue that after massive evaluations had been conducted, policymakers became skeptical of whether such studies could provide useful outcome and cost-benefit data on the overall effects of large educational interventions. At the same time, research conducted between 1965 and 1980 on the implementation of federal programs had yielded useful information about that process as it related to the distribution of program resources and the delivery of services to students. Not surprisingly, given their political implications, these Title I program outputs were the focus of policymakers' attention and expectations.

mation with three developments defining the changes: the focus of accountability has moved from monitoring the distribution of inputs to evaluating the effectiveness of program outcomes; states are now required to incorporate Title I recipients into their standards and assessment systems as they apply to all students; and the technical characteristics of tests have changed and their uses have become more consequential.

A Shift in Focus to Program Outcomes

The major changes in Title I's testing requirements began early in ESEA's third decade with its 1988 reauthorization.² States were required for the first time to define the levels of academic achievement that Title I eligible students should attain as a way to identify schools whose students did not show substantial progress in meeting the achievement outcomes (Jennings 2001, 15). This new focus represented a significant shift in Title I's rationale by highlighting the academic achievement of Title I students, and by beginning to identify them not just as recipients of special services but also as participants in a school's general academic program (Manna 2006, 73).³ Nevertheless, despite emphasizing accountability for program outcomes and articulating an explicit connection between Title I services and the general education program, the 1988 reauthorization still framed Title I recipients as a distinct group. That segregation was further reinforced because the required Title I testing regime functioned as a separate system that affected only those 20 percent of the nation's

students who were Title I eligible (Manna 2006, 75).⁴

In 1993, the Advisory Committee on Testing in Chapter 1 [Title I], established by the Department of Education (ED) to review the standardized tests used to evaluate the program, issued its report. The committee, composed primarily of testing experts and other researchers, concluded that "Chapter 1 testing should no longer be an independent system but should be linked with the education reforms that states and school districts are undertaking for all children," and that "national Chapter 1 evaluation should be decoupled from state, local, and classroom assessment functions" (vii). The report also noted that researchers and practitioners were finding that Chapter 1 procedures "may be narrowing Chapter 1 curriculum and instruction by rewarding those practices most likely to produce gains on norm-referenced tests," and that such tests may encourage teachers to spend too much time teaching low-level skills and test preparation (13).

Title I Linked to State Standards and Assessments

In the 1994 ESEA reauthorization, Improving America's Schools Act (IASA), the Clinton administration relied heavily on the testing provisions to cement the connection between Title I and the general education program in states and local districts. As a condition for receiving federal funding, states were required to ensure that the learning goals and standards for Title I students were the same as for all other students. Although Title I students might receive supplemental instruc-

2. Between 1981 and 1994, Title I was known as Chapter 1. With the 1994 reauthorization of ESEA (Improving America's Schools Act [IASA]), it was once again called Title I.

3. The Title I program and its students were moved closer to the core instructional program in individual schools with the initiation of school-wide programs. In response to concerns that pull-out services were preventing eligible students from receiving full exposure to the regular curriculum and to findings from the effective schools research showing that successful schools are characterized by comprehensive instructional strategies, the 1988 ESEA reauthorization allowed schools in which 75 percent of the students came from low-income backgrounds to operate school-wide programs without requesting matching funds from the local school district.

4. As states adopted standards-based reforms and enacted accountability systems that focused on student outcomes, state officials—particularly governors—pressed the federal government to move away from its policy that allowed and even encouraged tests of basic skills and to make the Title I testing provisions more consistent with state policy (McDermott 2011, 69).

tion, schools now had to ensure that these students were part of the core instructional program, and schools had to be accountable for the Title I students' academic progress in whatever way states held them accountable for all other students' achievement.

The assumptions underlying IASA and its testing requirements were those articulated in the academic writing of Clinton's undersecretary of education, Marshall Smith, and reflected in the policies of states that had adopted some form of standards-based reform (SBR) (Smith and O'Day 1991; O'Day and Smith 1993).⁵ Although the strategy has varied from one jurisdiction to another, four elements have typically characterized it: a focus on student achievement; an emphasis on academic content standards specifying the knowledge and skills that students should acquire and the levels at which they should demonstrate mastery; a desire to extend the standards to all students, including those for whom expectations had traditionally been low; and a heavy reliance on achievement testing to spur the reforms and monitor their impact (National Research Council 1997).

IASA required that state assessments had to be aligned with the content standards, test at three separate grade levels, be based on "multiple, up-to-date . . . measures that assess higher order thinking skills and understanding," and "provide individual student interpretive and descriptive reports" as well as disaggregated results at the school level by race, gender, English proficiency, migrant status, disability, and economic status (Improving America's Schools Act, section 1111). States were required to hold schools and districts accountable for making adequate progress toward achieving the standards, and they were

to identify districts and schools in need of improvement and to take corrective action in cases of persistent academic failure.

Congress gave the states a long implementation period, allowing them to implement major provisions of IASA over six years with the assessments not required to be aligned with a state's content and performance standards until the 2000–2001 school year. However, by early 2001, only seventeen states were prepared to meet the deadline for aligned assessments that tested all students in reading and mathematics at least once in the elementary, middle, and secondary grade spans.⁶ In addition, states varied considerably in how they implemented the required student performance standards. Although all but five had set absolute goals for student performance, they had significantly different expectations about the proportion of students who would need to meet the state's definition of proficiency. Twelve states expected 90 to 100 percent of students in each school to meet the state's proficiency standard, and another ten set a goal of 50 percent. Only fourteen states had specific time lines for meeting performance standards, on average ten years with a range of six to twenty years. States also used different methods for defining adequate yearly progress (AYP). Some required schools to meet an absolute performance target, others expected relative improvement each year or reductions in the achievement gap among subgroups of students, and still others used various combinations of these approaches. States also varied in the proportion of Title I schools they designated as "needing improvement," ranging from a low of 1 percent in Texas and 5 percent in North Carolina to a high of 76 percent in Michigan (Cohen 2002, 5).

5. The rationale for what they call systemic reform is that if states set high academic content and performance standards common to all students, the wide gaps in achievement among students of different ethnic and income groups would narrow. Their vision of reform is one where state governments set the standards, but allow individual districts and schools to decide on the instructional strategies to meet them. The systemic notion refers to the close links among curriculum, professional development, and school organization in implementing academic standards.

6. An additional fourteen states had been granted a waiver on the implementation deadline, but were expected to meet the requirements given some additional time. Four other states, Alabama, California, Wisconsin, and West Virginia, were found to have been substantially out of compliance and unlikely to meet the requirements without federal enforcement action (Cohen 2002, 3).

More Consequential Uses of Test Results

A number of scholars have analyzed the legislative history of No Child Left Behind (NCLB), IASA's successor (Rudalevige, 2003; DeBray 2006; Manna 2006; McGuinn 2006). Among the issues that they document as contentious were the testing provisions, stemming mainly from some states' opposition to changing their existing assessment systems. Once again, the federal government faced the dilemma of seeking greater accountability over how its funds are spent, but depending on the states for the necessary information. The challenges associated with the federal government continuing to depend on state assessments but being able only to set general guidelines for their design and administration were compounded by concerns about test use. Because AYP was viewed as both an indicator for reporting student achievement uniformly across all the state testing systems and a tool for leveraging state and local behavior, it was the focus of considerable debate. The requirement that all student subgroups within a school meet the AYP standard and that, for the first time, consequences would be tied to test scores in the form of potential sanctions meant that it had become a high-stakes indicator and policy tool. Before and after NCLB's passage, researchers warned that AYP and the consequences attached to its use compromised its validity as a measurement instrument (Kane and Staiger 2002; Linn, Baker, and Betebenner 2002). While the legislation was in the conference committee, congressional staff scrambled to revise the AYP formula even though researchers warned that it would eventually result in large numbers of failing schools (Manna 2006, 125). Neverthe-

less, NCLB passed both houses of Congress with large majorities, and with the testing provisions essentially intact from what the Bush administration had originally proposed (Manna 2006, 127).

The NCLB testing mandates represent a more precise and detailed version of the IASA requirements, and they embody stronger regulatory teeth in moving test use from essentially informational, hortatory uses to ones with high-stakes consequences.⁷ As a condition for receiving Title I funding, states are required to test all students in grades three through eight annually in mathematics and reading–language arts and once in grades nine through twelve on a statewide standardized assessment aligned with the state content standards.⁸ Test scores are to be disaggregated and reported at the school-, district-, and state-levels by race-ethnicity, gender, low income, disability, and students learning English. States must also participate in the fourth- and eighth-grade National Assessment of Educational Progress (NAEP). The high stakes aspect of NCLB's testing regime is represented in an AYP formula that requires states to set annual targets for increasing student achievement and closing gaps among groups so that by the 2013–2014 academic year, all students were to be proficient in mathematics and reading as measured on their state assessments. NCLB sanctions for schools failing to meet AYP include giving parents the option to transfer their children to other schools, using part of their Title I support to provide parents with funds to secure supplemental assistance for their children, and in cases of failing to meet AYP over four consecutive years, undergoing major restructuring with the possibility of the school closing

7. Low- and high-stakes tests represent two fundamentally different ways of using testing in service of policy goals. "A low-stakes test has no significant tangible, or direct consequences attached to the results, with information alone assumed to be a sufficient incentive for people to act. The theory behind this policy is that a standardized test can reliably and validly measure student achievement; that politicians, educators, parents, and the public will then act on the information generated by the test; and that actions based on test results will improve educational quality and student achievement. In contrast, high-stakes policies assume that information alone is insufficient to motivate educators to teach well and students to perform to high standards. Hence it is assumed, the promise of rewards or the threat of sanctions is needed to ensure change" (National Research Council 1999a, 35).

8. In addition to testing in mathematics and reading, states are also required to test students in science once in elementary, middle, and high school.

and being reorganized under new management. As a number of evaluations of NCLB have found, the results have been decidedly mixed (for a summary of these studies, see Dee and Jacob 2010). Nevertheless, even when the Obama administration offered waivers in the face of states' inability to meet the 2014 proficiency standard and Congress's failure to address problems with NCLB and reauthorize ESEA, the annual testing and reporting requirements remain in place.⁹ The waivers have offered states the opportunity to alter how they use their test scores for accountability purposes, but the frequency of testing and the subjects and grades tested are unchanged.

NCLB's mixed results are reflected in how its requirements have served the two purposes of testing. As policy tools, they have reinforced what was already known from state policies: mandates that students be tested on standardized assessments are the most powerful levers that elected officials and other policymakers have for influencing what happens in schools and classrooms. A growing body of research has found that although the changes may not have the desired or expected effects on student learning, they do in fact change school and classroom practices (National Research Council 1999a, 29). In complying with the testing requirements that NCLB imposed on them as a condition for Title I funding, states made major changes in their assessment systems. At the time of NCLB's enactment, forty-six administered a statewide assessment, and thirty-one reported that the tests were aligned with their

state standards. However, because most states tested at just a few grade levels or tested only a sample of students, only five fully met the NCLB requirements. The remainder had to develop at least one new test and thirty-five states had to develop seven or more new tests (U.S. General Accounting Office 2003, 8, 13).

At the same time, the AYP requirements, coupled with the provision that each state could develop its own content and performance standards, created incentives for some states to shirk by setting their standards low and making it easier for schools to reach AYP. As a result, Robert Linn, an expert on state assessments, concludes that "the variability in the stringency of the state standards defining proficient performance is so great that the concept of proficient achievement lacks meaning" (2008, 7). This discrepancy in standards became clear when state assessment results were compared with NAEP scores. In mapping state proficiency standards in mathematics and reading for grades four and eight onto the appropriate NAEP scale, researchers found that state differences in the percentage of students scoring at the proficient level on state assessments did not represent real differences in achievement as measured on NAEP, but instead reflected where a state set its proficiency levels. Most state cut points, moreover, fell below the equivalent of the NAEP proficient standard, and some even fell below the NAEP basic standard (National Center for Education Statistics 2007).¹⁰

The effects of NCLB's testing provisions in

9. The department first offered waivers to states in 2011. To be exempt from having to meet the 2014 proficiency standard and to implement specific interventions in schools that fail to meet AYP for two consecutive years, states could apply for a waiver subject to annual renewal. The waiver includes four requirements. First, they must adopt college- and career-ready standards and administer high-quality assessments based on those standards by the 2014–2015 school year (states were allowed to continue to use their earlier SBR assessments in the interim). Second, they must develop differential accountability systems that include multiple indicators in addition to test scores (for example, graduation rates, student attendance) and that are publicly reported by student subgroup (though the number of groups can be fewer than those required for NCLB if the categories are broader). Third, they must implement teacher and principal evaluation systems that factor in student achievement. Fourth, they must identify three levels of schools based on their performance (*reward*, *priority*, and *focus*) with interventions required for the lowest performing (priority—equal to at least 5 percent of the state's lowest performing schools) and those with large achievement gaps among subgroups (focus schools). As of the fall of 2014, the department had granted forty-three state waivers and had withdrawn two of these.

10. Comparing NAEP scores for a given state with those on the state's own assessment is not entirely valid. The NAEP tests are not aligned with the standards or curriculum of any particular state and the NAEP proficiency

functioning as a measurement instrument are also mixed. The costs of meeting the requirement to test all students in grades three through eight annually has led states to increase their use of multiple-choice items in their assessments because of the ease of scoring within the time frame specified by NCLB. In 2009, thirty-eight of the forty-eight states responding to a Government Accountability Office (GAO) survey, reported that multiple-choice items made up all or most of the items for their reading and language arts assessments and thirty-nine reported the same format for their mathematics assessments. About 20 percent of the states reported increasing their use of multiple-choice items since NCLB's passage. State officials and their technical advisers acknowledged the significant trade-offs they have faced between validly measuring cognitively complex content and accommodating cost and time constraints. The GAO found that some states have attempted to address these trade-offs by including open-ended, constructed response items on their assessments that are outside the NCLB reporting requirements and used only to provide information for instructional purposes (2009).

However, during this same period, NAEP has become a more visible and credible source of information about student achievement. The NCLB requirement that all states participate in NAEP and the introduction of the Trial Urban District Assessment (TUDA) in 2002 has provided more uniform data about student performance. In addition, it has highlighted the shortcomings of state performance standards and has functioned as the main source

of information about differential student outcomes during national policy debates.¹¹

From their beginnings in 1965, an overarching function of the Title I testing requirements has been their role in evaluating the effectiveness of the program. That task has not produced clear-cut answers, and in the case of NCLB, has been the subject of considerable contention. However, even researchers who have attributed positive changes to Title I agree that the effects have been modest. In their synthesis of seventeen federally commissioned evaluations of Title I's effectiveness between 1966 and 1992, Geoffrey Borman and Jerome D'Agostino found that the program "has not fulfilled its original expectation: to close the achievement gap between at-risk students and their more advantaged peers. . . . The results do suggest, however, that without the program, children served since the 1960s would have fallen further behind academically" (Borman and D'Agostino 2001, 49). These authors conclude, as have others, that although Title I has produced a modest effect on students' annual achievement gains, the effect has been highly variable across subject areas, testing cycles, grade levels, and schools. Subsequent studies of the effects of NCLB, using a variety of analytical techniques and primarily relying on NAEP and state assessment data, have reached differing conclusions. Some have found no achievement effects associated with NCLB, and others have found either that growth in student achievement has been flatter since the enactment of NCLB or that it tracks trends that existed prior to NCLB. In contrast, a more recent study by Thomas Dee

standards have been set very high so scores on that assessment can systematically overstate the shortcomings of state performance standards and assessments. However, another large data base from tests aligned with individual state standards confirms the NAEP results. The Northwest Evaluation Association (NWEA) maintains a large pool of test items from which it designs diagnostic tests for local districts using items closest to the relevant state assessment. Because all the items are pegged to a common scale, NWEA can make comparisons across the states. Using data from 2003 and 2006, collected on 830,000 students in twenty-six states, NWEA found significant variability in the level of difficulty in state assessments ranging from the 6th percentile (94 percent would pass) to approximately the 77th percentile (23 percent would pass) (National Research Council 2008).

11. Another development, implemented by states independently of the NCLB requirements, has also contributed to the design of more reliable and comprehensive student data systems and more effective use of them. Over the past decade, states have begun to develop longitudinal student-level data systems and to link their student and educator data, pre-K-12 and postsecondary systems, and across education, social service, and employment agencies.

and Brian Jacob that compares test score changes between 1990 and 2007 across states that already had school accountability policies before NCLB and ones that did not finds modest and mixed effects. They report that NCLB has been associated with statistically significant increases in the average mathematics performance of fourth graders with somewhat larger effects among the highest- and lowest-achieving students. The effects on eighth-grade mathematics scores are also positive, especially for low-achieving groups. However, there is no evidence that NCLB has had any effect on either fourth- or eighth-grade reading scores. The researchers note that these achievement gains appear limited when compared with NCLB's goal of 100 percent proficiency and that the program has contributed only modestly to reducing the achievement gap (Dee and Jacob 2011).

EXPLAINING STABILITY AND CHANGE IN TITLE I TESTING POLICY

The history of the Title I testing requirements raises two questions:

- Why in the face of continuing evidence about the limitations of testing as a policy tool and measurement instrument have these provisions endured through multiple ESEA reauthorizations?
- Why have significant changes been adopted even while the core elements of the testing requirements persist?

Perhaps surprisingly, the same two factors explain both the stability and changes: the institutional structures and rules that make federalism a defining feature of government in the United States, and the interest group dynamics that have shaped ESEA.

The Influence of Federalism

More than in some other policy domains, federal authority in education is essentially limited to enforcing constitutional civil rights and civil liberties guarantees. Beyond that, the federal government's major policy tools are categorical programs that seek to change the institutional behavior of state and local agencies by offering financial assistance on the condition that they undertake certain prescribed activities (McDermott and Jensen 2005; McDonnell 2005). This arrangement has two central features. First, because schools are "coping organizations" where neither outputs nor outcomes are truly observable, the federal government must rely on proxy indicators and limited information (Wilson, 1989, 168). Consequently, information about how Title I funds are spent has become the proxy for outputs and test scores the main proxy for outcomes. The fifty-year history of ESEA suggests that a wholly different strategy is not likely to be politically or administratively feasible.

The second feature is that the federal government's enforcement powers in the case of Title I are limited. In theory, a categorical program carries the threat of the withdrawal of funds if recipients fail to comply with the conditions of the grant. However, although the Department of Education has been willing to impose some partial withholding of Title I administrative funds, it has avoided more stringent penalties because of the likely political pushback from Congress and the potential harm to students receiving Title I services.¹² Consequently, the shaming of malefactors through information dissemination and bargaining with states have been the modal enforcement mechanisms.

Although federalism has functioned as an institutional constraint ensuring the stability

12. A recent example of the federal government's limits on enforcing its categorical requirements was ED's response to California's decision to suspend accountability testing and reporting required under NCLB for one year as it field-tested new Common Core assessments in 2014. The initial ED response was to threaten to withhold at least \$15 million in Title I administrative funds from California with the possibility that the amount would increase as the federal government also withheld ESEA funds spent on testing in the previous year and on turnaround schools. However, six months later after California's governor, legislature, state board of education, and state education agency refused to change their position, ED secretary Arne Duncan granted the state a waiver. California students took the pilot test in 2014, but the results will not be publicly reported by school and subgroup and will not be included in the formula for calculating AYP (Klein 2013; McNeil 2014).

of the testing requirements, it has also provided the federal government with opportunities for changes that have extended and strengthened its programmatic reach over state and local behavior. Several analysts have concluded that NCLB's design, especially the more prescriptive testing requirements, was possible only because of profound changes in the state role beginning in the 1980s (McDonnell 2005; Manna 2006). In the wake of the "Nation at Risk" report and the implementation of standards-based reforms in a number of states, academic content and performance standards along with standardized assessments became an integral part of state policy (McDermott 2011). This major development at the level of the governmental system with constitutional responsibility for education allowed the federal government "to borrow strength." Paul Manna defines this process as occurring "when policy entrepreneurs at one level of government attempt to push their agendas by leveraging the justification and capabilities that other governments in the federal system possess" (2006, 5). Policy entrepreneurs promoting NCLB could mobilize around the license, or arguments that states had already made to justify the involvement of higher levels of government in classroom processes and outcomes, and around the capacity or resources and administrative structures that state reforms had created.¹³

The last point about capacity and structures is particularly important for the Title I testing requirements. One effect of more than twenty years of state SBR policies is that a substantial testing infrastructure has been institutionalized (McDonnell 2008). Networks of state agency staff, testing contractors, vendors of instructional materials, and local testing and evaluation staff are now well developed. Most

states had to change their testing policies in response to NCLB, but the institutional infrastructure was already firmly in place. That capability and the policy ideas animating it have allowed substantial changes in Title I testing over the past two decades yet ensured that its core elements remain stable.

The Role of Interest Groups

A number of recent studies have analyzed the changing politics of education through the lens of groups with a stake in this policy domain. Taking somewhat different but complementary perspectives, this research has focused on the growing density and diversity of interest networks and on the altered issue definitions and policy ideas that the groups have embraced (for example, see DeBray-Pelot and McGuinn 2009; Rhodes 2012; Mehta 2013; Wolbrecht and Hartney 2014). These interest group dynamics—their ideological and material interests, how they frame them, and their policy preferences and strategies—are a second factor in explaining Title I's change within stability. This interest-focused explanation also has institutional dimensions that closely connect it to the first explanation. Just as the federal government borrowed strength from the states, groups with an interest in testing as part of a reform strategy have taken advantage of the multiple policy arenas in the federal system in advancing their agenda. Their promotion of SBR at all governmental levels has resulted not only in a range of policies, but also in new institutions to develop and maintain those reforms. The testing infrastructures now operating in states and local districts and used in implementing the NCLB requirements are prime examples.¹⁴

One of the most noteworthy changes in the politics of education since the 1980s has been

13. Although borrowing strength is typically used to explain the emergence of new federal policies that are legitimated through existing state ones, the process can also work in the opposite direction. So, for example, although state policymakers may oppose the strictures imposed by NCLB, the law has allowed them to use the federal mandate as leverage to strengthen their own influence over local districts and schools.

14. Jennifer Hochschild notes that theories of issue expansion identify the creation of new institutions as a strategy that interest groups use to maintain their policy gains even as public attention shifts to other problems and opponents seek to diminish or alter those gains. Once established, institutional rules and incentives are more difficult to change than policies, and Hochschild suggests that this difference helps account for the continuing success of groups pressing for greater educational accountability (2003, 119–20).

a substantial expansion in the array of groups beyond those representing the interests of teachers, administrators, and education governing bodies. Although the traditional education interest groups disagree among themselves on a number of issues such as labor relations, they have historically coalesced in pressing for greater state and local flexibility in administering Title I, and they have opposed the NCLB testing and AYP requirements as overly prescriptive. In contrast, newer entrants into the education policy arena, such as business interests and groups promoting various reform agendas, actively support the NCLB testing requirements as part of a larger accountability strategy. Their support has ensured the continuation of testing as a policy tool central to Title I, but it has also helped promote the changes that have made the requirements more prescriptive.

The reasons for these groups' support can be traced to a combination of revised problem definitions and new policy ideas framed as solutions to those problems. Because different groups have accepted and promoted alternative problem definitions and rationales, their shared agreement on stricter accountability provisions as a solution does not always extend to the reasons for it. Space limitations preclude a thorough discussion of the new problem framings and policy ideas that contributed to changes in the Title I testing requirements. However, several of these are specific to ESEA, and others reflect broader changes in education politics and policy. One factor, directly linked to ESEA's history and reflected in the congressional deliberations over NCLB, was a perception on the part of much of the Republican caucus and many moderate Democrats that federal education policy had not demanded real results for the billions of dollars spent (Rudalevige 2003; DeBray 2006). Consequently, NCLB was seen as a way to deal with the persistent and vexing problem of Title I's modest effects by moving federal regulation away from an emphasis on fiscal audit trails,

and trading some increased flexibility in program operations for states and localities in exchange for their greater accountability for student outcomes. Frustration with Title I's shortcomings also motivated some civil rights groups who view SBR with a strong testing and accountability component as a way to focus attention on the underachievement of historically disadvantaged students, and to create political momentum for improving the schools they attend. These groups, like Robert Kennedy decades earlier, believe that states and localities will not adequately serve disadvantaged students without federal pressure to do so.¹⁵

Title I's own history and how past evaluations have shaped definitions of the policy problem partly explain changes in the testing provisions. However, the selection of a standards-based strategy with high-stakes assessment as a central feature of the solution is best explained by the broader SBR rationale. That rationale is now well known, having been repeated in media commentary and policy deliberations dating back to the publication of "A Nation at Risk" in 1982. Its statement of the problem includes an economic and demographic dimension: to be competitive in a technologically advanced, global economy, the United States needs better-educated workers, and the changing demographics of the U.S. labor force require that schools do a more effective job of educating those students who have historically been poorly served by the public schools. Much has been written about how this problem definition became linked to standards and accountability and school choice as the solutions now dominating education policy (for example, see Manna 2006; Rhodes 2012; Henig 2013). Similarly, numerous researchers and commentators have questioned the underlying assumptions of these policy ideas and the evidence about their effectiveness (for example, see Cohen and Moffitt 2009; Ravitch 2010; Kirp 2013).

What is perhaps most important from the

15. In his analysis of what he calls "civil rights [policy] entrepreneurs," Jesse Rhodes makes a distinction between these groups as one faction within the civil rights community and other organizations that do not support SBR. The groups working on behalf of NCLB and related state policies include the Citizens' Commission on Civil Rights, the Education Trust, and the National Council of La Raza.

perspective of Title I is that these policy ideas moved the program and its recipients from the periphery of schooling to the instructional core with its academic standards and accountability requirements. This shift occurred because the dominant policy image of what constitutes educational equity was changed. A problem statement requiring that all students be educated to higher standards shifted the definition of equity away from access to educational resources and compliance with legal mandates to a focus on students' learning opportunities and their achievement. As Kathryn McDermott concludes, "by defining equity in terms of a common educational threshold for all students, the performance-based understanding of educational equity shifts to a universal definition of equity and away from understandings of equity that targeted specific disadvantaged groups such as low-income students, students of color, or girls" (2011, 167). As a result, Title I, in effect, became an example of what Theda Skocpol calls antipoverty programs based on "targeting within universalism" (1991). In this case, the testing requirements have placed Title I within a universal policy framework, and the interests promoting the broader SBR idea have made that shift politically feasible.

CONCLUSION

Three conclusions emerge from the history of the Title I testing and evaluation requirements: their dual policy and educational purposes, the persistence of change within a stable policy core requiring regular student testing, and the institutional and interest-based reasons for how the testing requirements have evolved. Looking ahead to ESEA's future, questions arise about whether the testing requirements will survive and, if so, will they continue to serve the same two purposes. Because a number of groups and commentators are calling attention to problems with the overtesting of students and poor quality assessments, the testing requirements will likely become less prescriptive in a reauthorized Title I. Nevertheless, despite significant disagreements between Democratic and Republican members of Congress on major parts of the legislation, the requirement that students be tested ap-

pears in most proposals, the leadership still focused on students continuing to be tested in grades three through eight, and some rank-and-file members pressing for less testing but still requiring it on a regular basis (Camera 2014; Rich 2015a, 2015b).

What is likely to change is which governmental level establishes the rules for state accountability plans and how test results are used in rewarding and sanctioning schools and educators. If states are given greater flexibility in how often they test students, some may decide that with less frequent testing (or testing only samples of students), they can afford to increase the validity of their assessments through improved item design and curriculum coverage. Increased state flexibility in test use may also move assessments back to earlier low-stakes, hortatory uses that depend on transparency and reducing information asymmetries among policymakers, their constituents, and educators. If granting states greater flexibility leads them to administering assessments that more validly measure student performance and using the scores only for purposes for which a test is designed, the result will be higher quality assessment systems more closely aligned with the standards established by the testing profession (American Educational Research Association et al. 2014). However, not all states will use such flexibility to improve their tests as either policy tools or measurement instruments.

Consequently, if the federal government is to maintain its core policy goal of enhanced learning opportunities for low-income students, two key elements of NCLB need to remain in a reauthorized ESEA. The requirement for reporting the distribution of test scores by student subgroups has been one of the most effective examples of the federal bully pulpit in highlighting social problems and in providing a major resource for political mobilization. Similarly, the requirement that states participate in NAEP has generated substantial pay-offs for education reformers as comparisons between state assessment results and NAEP scores (as flawed as these comparisons may be) have functioned as powerful rationales for subsequent policies such as the Common Core

State Standards. So a stable core of required testing and public reporting of the results should continue in a reauthorized ESEA even if the level of state discretion over test design and use is increased.

If such changes occur, the issue will be whether testing can continue to function as a policy tool and measurement instrument. Despite continued criticisms and identified flaws, the testing requirements have proven to be among the federal government's most effective strategies for pursuing its ESEA social policy goals. As a tool for leveraging state and local behavior, the requirements have significantly increased the federal government's influence over the allocation of program resources and its ability to reach further down into the education system in shaping decisions about who will be served with Title I funds and how they will be served. The evaluations conducted during Title I's early years led to more precise targeting of program resources, and the SBR orientation reflected in IASA and NCLB integrated Title I recipients into the instructional core with other students. Consequently, the testing requirements will continue to function as a policy tool in providing the information that allows federal policymakers to know whether subnational governments are meeting the categorical conditions for receiving funding, to negotiate with states from a stronger bargaining position, and to provide constituents with a mobilization resource.

The record of Title I testing requirements as a measurement instrument has been mixed. The requirements led to expanded use of student testing and increased capacity of states, local districts, and commercial test developers in designing and using standardized assessments. At the same time, Title I has also created disincentives for developing more reliable and valid tests. Most testing experts would argue that the move from norm-referenced tests to standards-based ones improved the validity of the inferences that could be drawn about student progress because they were measuring knowledge and skills more closely aligned with what was being taught in individual states. However, these advances were halted by the cost and administrative constraints imposed by the NCLB testing re-

quirements and the move to more multiple choice testing. How the testing requirements continue to function as a measurement instrument will depend on the amount of flexibility that states (either individually or in consortia) will have in designing their assessments and on how the federal government will decide to incorporate the use of test results in ESEA program rules and administration. What can be predicted with some certainty, however, is that the status of the measurement function will continue to depend on the demands placed on testing when it is used as a political and policy tool.

The degree to which federalism is embedded in the structure and political culture of the United States has created a symbiotic relationship between the policy and measurement purposes of the Title I testing requirements, but it has also generated some unproductive trade-offs. On the one hand, pursuit of the federal government's policy purposes has politicized Title I testing and its uses in ways that have weakened its quality as a measurement instrument, especially to the extent that it has created incentives for educators to narrow their teaching to the content being measured (National Research Council 2011). At the same time, the federal government has been limited in how it can use testing for policy purposes because it must depend on the states to be its enforcer, information source, and main policy implementer.

Consequently, the Title I testing requirements have never completely fulfilled Robert Kennedy's vision of objective measurement in the service of effective programs for students living in poverty. Yet when we consider the institutional constraints on federal action in education and the competing demands placed on Title I by an expanding range of interests, the testing requirements have been more successful and enduring than what might have been predicted early in ESEA's history. The challenge now is the same one the program has faced for fifty years: ensuring state and local behavior consistent with federal goals while acknowledging the technical limitations of testing and balancing political accountability by elected officials with educators' professional judgment.

REFERENCES

- Advisory Committee on Testing in Chapter 1. 1993. *Reinforcing the Promise, Reforming the Paradigm*. Washington: U.S. Department of Education.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Bailey, Stephen K., and Edith K. Mosher. 1968. *ESEA: The Office of Education Administers a Law*. Syracuse, N.Y.: Syracuse University Press.
- Borman, Geoffrey D., and Jerome V. D'Agostino. 2001. "Title I and Student Achievement: A Quantitative Synthesis." In *Title I: Compensatory Education at the Crossroads*, edited by Geoffrey D. Borman, Samuel C. Stringfield, and Robert E. Slavin. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Camera, Lauren. 2014. "AFT Backs Newest Proposal to Reduce Testing." *Education Week*, September. Available at: http://blogs.edweek.org/edweek/campaign-1-12/2014/09/aft_backs_n (accessed October 6, 2014).
- Cohen, Michael. 2002. "Implementing Title I Standards, Assessments, and Accountability: Lessons from the Past Challenges for the Future." Paper prepared for the conference Will No Child Be Left Behind? The Challenges of Making This Law Work, Washington (February 13, 2002).
- Cohen, David K., and Susan L. Moffitt. 2009. *The Ordeal of Equality*. Cambridge, Mass.: Harvard University Press.
- DeBray, Elizabeth H. 2006. *Politics, Ideology, and Education*. New York: Teachers College Press.
- DeBray-Pelot, Elizabeth. 2007. "Dismantling Education's 'Iron Triangle.'" In *To Educate a Nation*, edited by Carl F. Kaestle and Alyssa E. Lodewick. Lawrence: University Press of Kansas.
- DeBray-Pelot, Elizabeth, and Patrick McGuinn. 2009. "The New Politics of Education: Analyzing the Federal Education Policy Landscape in the Post-NCLB Era." *Educational Policy*, 23(1): 15–42.
- Dee, Thomas S., and Brian Jacob. 2010. "The Impact of No Child Left Behind on Students, Teachers, and Schools." *Brookings Papers on Economic Activity* Fall (2010): 149–94.
- . 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30(3): 418–46.
- Fung, Archon, Mary Graham, and David Weil. 2007. *Full Disclosure*. New York: Cambridge University Press.
- Graham, Hugh David. 1984. *The Uncertain Triumph*. Chapel Hill, N.C.: University of North Carolina Press.
- Henig, Jeffrey. 2013. *The End of Exceptionalism in American Education*. Cambridge, Mass.: Harvard Education Press.
- Hochschild, Jennifer. 2003. "Rethinking Accountability Politics." In *No Child Left Behind*, edited by Paul E. Peterson and Martin R. West. Washington, D.C.: Brookings Institution Press.
- Ingram, Helen. 1977. "Policy Implementation Through Bargaining." *Public Policy* 25(4): 499–526.
- Jennings, John F. 2001. "Title I: Its Legislative History and Its Promise." In *Title I: Compensatory Education at the Crossroads*, edited by Geoffrey D. Borman, Samuel C. Stringfield, and Robert E. Slavin. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Kaestle, Carl F., and Marshall S. Smith. 1982. "The Federal Role in Elementary and Secondary Education, 1940–1980." *Harvard Educational Review* 52(4): 384–408.
- Kane, Thomas J., and Douglas O. Staiger. 2002. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." *Brookings Papers on Education Policy* 5: 235–83.
- Kirp, David L. 2013. *Improbable Scholars*. New York: Oxford University Press.
- Kirst, Michael, and Richard Jung. 1991. "The Utility of a Longitudinal Approach in Assessing Implementation: A Thirteen-Year View of Title I, ESEA." In *Education Policy Implementation*, edited by Allan R. Odden. Albany: State University of New York Press.
- Klein, Alyson. 2013. "Ed. Dept.: Calif. Could Lose at Least \$15 Million in Federal Funds Over Testing." *Education Week*, October. Available at: <http://blogs.edweek.org/edweek/campaign-k-12/2013/10/education> (accessed September 30, 2014).
- Linn, Robert L. 2008. "Educational Accountability Systems." In *The Future of Test-Based Educational Accountability*, edited by Katherine E. Ryan and Lorrie A. Shepard. New York: Routledge.
- Linn, Robert L., Eva L. Baker, and Damian W. Betebenner. 2002. "Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001." *Educational Researcher* 31(6): 3–16.

- Manna, Paul. 2006. *School's In*. Washington, D.C.: Georgetown University Press.
- McDermott, Kathryn A. 2011. *High-Stakes Reform*. Washington, D.C.: Georgetown University Press.
- McDermott, Kathryn A., and Laura S. Jensen. 2005. "Dubious Sovereignty: Federal Conditions of Aid and the No Child Left Behind Act." *Peabody Journal of Education* 80(2): 39–56.
- McDonnell, Lorraine M. 2004. *Politics, Persuasion, and Educational Testing*. Cambridge, Mass.: Harvard University Press.
- . 2005. "No Child Left Behind and the Federal Role in Education: Evolution or Revolution?" *Peabody Journal of Education* 80(2): 19–38.
- . 2008. "The Politics of Educational Accountability: Can the Clock Be Turned Back?" In *The Future of Test-Based Educational Accountability*, edited by Katherine E. Ryan and Lorrie A. Shepard. New York: Routledge.
- McGuinn, Patrick J. 2006. *No Child Left Behind and the Transformation of Federal Education Policy, 1965–2005*. Lawrence: University Press of Kansas.
- McLaughlin, Milbrey Wallin. 1975. *Evaluation and Reform*. Cambridge, Mass.: Ballinger.
- McNeil, Michele. 2014. "California Wins Prized NCLB Testing Waiver." *Education Week*, March 7. Available at: <http://blogs.edweek.org/edweek/campaign-k-12/2014/03/california> (accessed September 30, 2014).
- Mehta, Jal. 2013. "How Paradigms Create Politics: The Transformation of American Educational Policy, 1980–2001." *American Educational Research Journal* 50(2): 285–324.
- Murphy, Jerome T. 1991. "Title I of ESEA: The Politics of Implementing Federal Education Reform." In *Education Policy Implementation*, edited by Allan R. Odden. Albany, NY: State University of New York Press.
- National Center for Education Statistics. 2007. *Mapping 2005 State Proficiency Standards onto the NAEP Scales*. Washington: U.S. Department of Education.
- National Research Council. 1997. *Educating One and All*. Committee on Goals 2000 and the Inclusion of Students with Disabilities. Edited by Lorraine M. McDonnell, Margaret J. McLaughlin, and Patricia Morison. Washington, D.C.: National Academies Press.
- . 1999a. *High Stakes*. Committee on Appropriate Test Use. Edited by Jay P. Heubert and Robert M. Hauser. Washington, D.C.: National Academies Press.
- . 1999b. *Testing, Teaching and Learning*. Committee on Title I Testing and Assessment. Edited by Richard F. Elmore and Robert Rothman. Washington, D.C.: National Academies Press.
- . 2008. *Common Standards for K-12 Education? Considering the Evidence*. Summary of a Workshop Series. Alexandria Beatty, rapporteur. Committee on State Standards in Education. Washington, D.C.: National Academies Press.
- . 2011. *Incentives and Test-Based Accountability in Education*. Committee on Incentives and Test-Based Accountability in Public Education. Edited by Michael Hout and Stuart W. Elliott. Washington, D.C.: National Academies Press.
- O'Day, Jennifer A., and Marshall S. Smith. 1993. "Systemic Reform and Educational Opportunity." In *Designing Coherent Education Policy*, edited by Susan Fuhrman. San Francisco: Jossey-Bass.
- Ravitch, Diane. 2010. *The Death and Life of the Great American School System*. New York: Basic Books.
- Rhodes, Jesse H. 2012. *An Education in Politics*. Ithaca, N.Y.: Cornell University Press.
- Rich, Motoko. 2015a. "White House Still Backs Annual Testing in Schools." *New York Times*, January 12. Available at: <http://www.nytimes.com/2015/01/13/education/arne-duncan-says-administration-is-committed-to-testing.html> (accessed July 29, 2015).
- . 2015b. "Reviewing Federal Education Law, Senator Seeks More Local Control." *New York Times*, January 14. Available at: <http://www.nytimes.com/2015/01/14/us/reviewing-federal-education-law-senator-seeks-more-local-control.html> (accessed July 29, 2015).
- Rudalevige, Andrew. 2003. "No Child Left Behind: Forging a Congressional Compromise." In *No Child Left Behind*, edited by Paul E. Peterson and Martin R. West. Washington, D.C.: Brookings Institution Press.
- Shepard, Lorrie. 2008. "A Brief History of Accountability Testing." In *The Future of Test-Based Educational Accountability*, edited by Katherine E. Ryan and Lorrie A. Shepard. New York: Routledge.
- Skocpol, Theda. 1991. "Targeting Within Universalism: Politically Viable Policies to Combat Poverty in the United States." In *The Urban Underclass*, edited by Christopher Jencks and Paul E. Peter-

- son. Washington, D.C.: Brookings Institution Press.
- Smith, Marshall, and Jennifer O'Day. 1991. "Systemic School Reform." In *The Politics of Curriculum and Testing*, edited by Susan Fuhrman and Betty Malen. Philadelphia, Pa.: Falmer.
- Stone, Deborah. 2012. *Policy Paradox*. New York: W. W. Norton.
- U.S. Congress. Office of Technology Assessment. 1992. *Testing in American Schools*. OTA-SET-519. Washington: Government Printing Office.
- U.S. General Accounting Office. 2003. "Title I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help States Realize Efficiencies." GAO-03-389. Washington: Government Printing Office. Available at: <http://www.gao.gov/new.items/d03389.pdf> (accessed July 30, 2015).
- U.S. Government Accountability Office. 2009. "No Child Left Behind: Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments." GAO 09-911. Washington: Government Printing Office. Available at: <http://www.gao.gov/assets/300/295827.pdf> (accessed July 30, 2015).
- Vinovskis, Maris A. 1999. "Do Federal Compensatory Education Programs Really Work? A Brief Historical Analysis of Title I and Head Start." *American Journal of Education* 107: 187-209.
- Wilson, James Q. 1989. *Bureaucracy*. New York: Basic Books.
- Wolbrecht, Christina, and Michael T. Hartney. 2014. "Ideas about Interests: Explaining the Changing Partisan Politics of Education." *Perspectives on Politics* 12(3): 603-30.