The Role of Assessment in Federal Education Programs

W. James Popham*

University of California, Los Angeles

Paper commissioned by the Center on Education Policy, Washington, D.C.
For its project on Rethinking the Federal Role in Education

November 2008

The views expressed in this paper are those of the author.

I am indebted to the following colleagues who were kind enough to respond to my request for assistance during the early days of organizing this analysis: Paul Barton, Christopher Cross, Gene Glass, David Grissmer, Lyle Jones, Carl Kaestle, Wayne Riddle, Robert Rath, Sue Rigney, Edward Roeber, Bella Rosenberg, and Roy Truby. I appreciate their remarkably helpful recommendations. Several of these individuals were also kind enough to review an early draft of the paper. Clearly, any shortcomings in this paper are mine, not theirs. I am also indebted to Kay Uchiyama, who took the lead in scrutinizing 50 years' worth of federal legislation related to U.S. educational assessment. Her efforts were truly invaluable.

EXECUTIVE SUMMARY

Recognizing the potential of accountability tests to alter classroom instruction, a brief recapitulation is given of the federal government's past influence on educational assessments. During the past 50 years, we have seen the function of federally engendered educational assessments shift from monitoring the use of federal funds for programs prescribed for statute-specified student populations to, instead, assuring the academic achievement of all students.

Federal sway over state-level accountability testing reached its zenith recently because of key federal legislation enacted shortly after the turn of the century. Against this backdrop, an attempt is made to provide a serviceable framework for rethinking what should be an appropriate federal role in U.S. educational testing. It is argued that two dominant questions must first be answered, namely, (1) What level of control should the federal government have over educational accountability tests? and (2) What should be the measurement mission(s) of those tests? Five control options for accountability tests, ranging from zero federal control to total federal control, are then presented. Next, consideration is given to three design dimensions that will govern the degree to which an accountability test is apt to have a beneficial impact on instruction, accountability, or curriculum.

It is concluded that, especially through the various reauthorizations of the Elementary and Secondary Education Act of 1965, not only has the federal influence on the nature of U.S. accountability tests markedly intensified during the last half-century, but the range of students affected by these tests has also expanded dramatically. Although federal education-related laws—the laws themselves—have definitely had an impact on both instructional practices and the curriculum, thus far the accountability tests spawned by federal laws—the tests themselves—have been designed only to support accountability functions, not instructional or curricular initiatives. It is argued that two issues, namely, an appropriate level of federal control and the proper measurement missions for important accountability tests, should frame any serious rethinking of the federal role in educational assessment. A concluding postscript reveals the writer's position regarding both issues.

INTRODUCTION

Test scores trump everything—or so it must seem to today's American educators. If a school's students score well on a state's annual accountability tests, just about everyone thinks the school is successful. Great test scores are seen to signify great schooling. Conversely, low test scores send the message that a school has seriously stumbled. To illustrate, even if a school's teachers are properly trained, its students highly motivated, and those students' parents swarm to a school in support of the school's efforts, when the school's students score poorly on their state's annual accountability tests, the school is regarded as a loser. Yes, test scores definitely seem to trump everything.

This "test-scores-trump" mentality may not be warranted, but such a view of how we should judge our schools is certainly pervasive. Accordingly, given this widely held perception, it is not surprising that whatever is assessed by significant educational tests will have a substantial impact on what transpires in America's schools. The more important the consequences

associated with an educational test's use are, the more profound will be the impact of that test on day-to-day classroom happenings.

The following analysis is intended to address the way in which education-related federal legislation might beneficially influence the nature of educational assessments and, as a consequence, what goes on in our schools. This paper is one of several such analyses focused on rethinking the federal role in U.S. education. Because a companion analysis in this series of papers (Hamilton, Stecher, and Yuan, 2008) addresses the topic of curricular standards, I will *not* be dealing with curricular issues in the following pages. It is apparent that educational tests, especially those measuring students' achievement, will be (and should be) influenced by the nature of the curricular outcomes being assessed. Nonetheless, more than a half-century of personal experience in public education persuades me that when the influence of curricular aspirations on education is compared with the impact of test results, then test results almost always triumph. Curricular standards often become little more than high-sounding rhetoric. It is when we actually *measure* students' attainment of curricular standards that we influence what goes on in schools.

Defining Terms

To avoid confusion, a bit of term-defining is in order. First off, I will be using the terms "tests" and "assessments" interchangeably. Because the term *test* often causes people to think only of traditional paper-and-pencil ways of measuring students (such as when teachers use multiple-choice or essay exams), many educators currently prefer to use the label *assessments* to characterize a wider range of techniques for measuring students' status—for instance, portfolio assessments, observational procedures, and performance tests. Fundamentally, though, both labels are pretty much the same. We use students' *overt* responses to these measurement tools (whether they're referred to as "tests" or as "assessments") in order to arrive at inferences about students' *covert* knowledge and skills. We oblige our students to make observable responses to educational tests for the simple reason that we can't see what's going on inside those students' skulls.

Tests, that is, assessments, become *high-stakes* when there are important contingencies linked to students' performances. Those contingencies include what happens to individual test-takers, such as when low-scoring students are denied grade-to-grade promotions or don't receive high school diplomas. However, when a school's test scores are used to evaluate a school's instructional success, this is also a contingency of considerable import to the teachers and administrators who staff that school. For example, when the instructional quality of today's public schools is dominantly determined by how many students reach "proficient" (or better) scores on state tests required by the No Child Left Behind Act (NCLB), those tests are unmistakable examples of high-stakes tests. This is true even though, in many states, there are no consequences for individual test-takers based on NCLB test results. Accordingly, any assessments that are supposed to function as "accountability" tests are almost certain to be regarded—accurately—as high-stakes tests.

Many of the tests used in our schools, of course, should not be characterized as high-stakes assessments. When classroom tests are employed by teachers to grade their students or, perhaps, to help teachers make adjustments in their ongoing instructional activities, such tests may well be low-stakes or even no-stakes assessments. Such classroom tests, however, as part of what's now referred to as the "formative assessment" process, can play a potent role in improving students' learning (Popham, 2008b).

Our federal government could conceivably become involved in various kinds of high-stakes tests (such as college entrance exams) or low-stakes tests (such as classroom quizzes). Yet, by all odds, the most likely federal role in the nation's educational assessments will revolve around *accountability tests.* This is because we use accountability tests to evaluate the caliber of the schooling that's being provided to our children, and this evaluative function is a societal mission best served by various levels of government. Accordingly, the following analysis will deal exclusively with *accountability* assessments. I do not minimize the role of assessments in evaluating educational programs or, in the case of formative assessment, for improving classroom instruction. However, these important applications of educational assessment will have to be dealt with by others.

Certain accountability tests, of course, are more significant than others, chiefly because more important evaluation-based decisions are riding on their results. For instance, if Accountability Test A's results are used only to rank schools in a local newspaper, while Accountability Test B's results might be used to restaff or close down a low-performing school, it is apparent which of these two tests will occupy a higher rung of the high-stakes assessment ladder.

Because educators (as is true with all of us) want to succeed, and because students' high scores on accountability tests are widely seen as educational "success," whatever is measured by accountability tests is certain to be emphasized by teachers in their classrooms. This is why the nature of our federal government's role in the nation's educational *testing* is so immensely important.

In preparing to undertake this analysis, I did some background reading regarding the impact federal legislation has had during the past half-century on the kinds of educational tests used in the United States. What I found was that substantial attention has been given to educational goals, that is, the knowledge or skills that we want our students to achieve. (I prefer to describe these goals as *curricular aims.*)

Since the 1980s, educational goals have often been referred to as "standards," or as "content standards." However, irrespective of what labels are used, far more federal attention has been given to educational goals than to the tests that will measure students' achievement of those goals. Many federal legislative battles, some surrounded by acrimony, have revolved around the nature and importance of educational standards. Yet, standards, or curricular aims, are often little more than educational yearnings on paper. Many sets of educational content standards, in truth, are merely curricular "wish lists" representing lofty aspirations about what is *hoped* students will somehow accomplish. What makes a genuine difference in schools is not what policymakers *say* is important for students to learn, but how those aspirations *are actually measured.* That's because test scores are linked to significant consequences. This is what inclines teachers to teach Topic *X* instead of Topic *Y.*

Federal Means of Influencing the Nation's Educational Assessments

Most readers of this analysis will already be conversant with the procedures our federal government can use to influence the nature of U.S. educational assessments. However, because some readers may not be, and because any rethinking of federal impact on educational assessment will surely involve new or modified federal legislation, a brief description of those procedures is in order. (Readers who are already conversant with such procedures might profitably skip ahead to page 8.)

Other than our Constitutional guarantee of fair and equal treatment for all U.S. citizens, federal influence on the nation's educational tests stems exclusively from federal statutes, that is, from federal laws which, directly or indirectly, address the properties or uses of any statute-specified educational assessments. However, because federal laws rarely explicate the totality of what is associated with a given law's intended outcomes, several additional layers of interpretive implementation are typically employed.

First off, federally enacted *statutes* are followed by *regulations* that, when approved, carry the same force of law as the provisions embodied in the originating law. These regulations, prior to their approval, are often the focus of substantial scrutiny. Sometimes public review of these regulations is called for in the statute itself because a law might require, for instance, the use of a "negotiated rule-making process." The regulations typically revolve around specifics about which the original law was silent or, in some instances, was ambiguous. (A law's regulations, of course, cannot contravene any provisions in the law itself.) At some point, however, the regulations associated with a federal law are ultimately approved. And if a federal law itself or its regulations deal with the nature and uses of state or local accountability assessments, then we are almost certain to see those federally identified requirements subsequently shape the nature and uses of the nation's accountability assessments.

After the statute and its accompanying regulations have been promulgated, federal personnel frequently issue one or more *guidance* documents. A guidance, as its name suggests, is intended to help the citizens most affected by a law comply with the law's provisions. A guidance is, therefore, a form of compliance advice so those who are supposed to follow a given federal law (and its related regulations) will know how to successfully do so. In relation to NCLB, for example, several guidance documents regarding educational assessments have been issued to those educators most concerned with the implementation of that law.

Finally, based on certain guidance documents, more elaborate procedures for supplying compliance advice are sometimes established. Regarding NCLB assessment, for instance, federal officials created a *peer-review process* whereby a series of independently appointed panels evaluate the degree to which each state's accountability tests are in accord with relevant NCLB assessment requirements. Similarly, on occasions the Secretary of Education has issued a *policy letter* to help educators see how best to comply with a federal law and/or its regulations. These sorts of mechanisms are, in essence, additional federal ways of helping relevant parties comply with the provisions of a federal law.

The peer-review process was a particularly powerful, potent way that federal personnel influenced states to satisfy NCLB's assessment requirements. Subsequent to the law's enactment, states were required to submit applications to show the U.S. Department of Education how they intended to satisfy the accountability and assessment provisions of NCLB and its subsequent regulations. If a state's application was not approved, then considerable federal dollars could be withheld from that state. Federal authorities used the expressed preferences of the peer-review panelists to move states toward what were seen to be more suitable ways of satisfying NCLB requirements. My personal conversations with numerous state assessment directors convince me that these peer-review pressures were remarkably influential in modifying states' NCLB accountability and assessment programs. To illustrate, the manner in which states now assess their students with disabilities is markedly different than it was prior to the enactment of NCLB and the influence of peer reviewers.

The NCLB peer-review process is clearly indicative of how much impact the federal government can have on educational tests if it wishes to do so. Even though this peer-review process seems

to represent a different category of federal influence on states' assessment-related decisions, it can still be technically classified as a form of NCLB-related *guidance* to state assessment personnel. Yet, given the gravity of the potential lost dollars if state authorities fail to heed the suggestions of the NCLB peer reviewers, the process might be more aptly characterized as federal "guidance with gusto."

Because the particulars of a federal statute obviously have the most significant impact on the kind of compliance advice needed thereafter, it is important to recognize that whereas some federal laws are remarkably circumscribing, other federal laws are much less constraining. To the degree that a federal statute is silent or ambiguous regarding certain features, the subsequent regulation and guidance activities of federal personnel may become not only more substantial, but also more influential. To illustrate, the assessment of English language learners has, in recent years, been influenced considerably by the way federal authorities have begun to interpret and implement certain features of NCLB. The more detailed and explicit a federal statute is, the less need there is for follow-on advice from federal officials.

It is surely debatable as to whether federal statutes *should* be more or less constraining. If a federal law is clear about its general intent, but vague about key features of its implementation, then this obliges federal personnel to help clarify the law's requirements. A lack of statutory constraints permits federal agencies to make along-the-way, experience-informed adjustments about how best to achieve the law's intent. To illustrate, it is reported that after the enactment of NCLB in early 2002, there were heated discussions among federal personnel regarding the law's "true" intent. Although there were records of the discussions that took place when the Congressional Conference Committee resolved differences between the House and Senate versions of NCLB, these "conference notes" captured only some—but not all—of those interpretive discussions. When a law's language is ambiguous, then, it may be literally impossible to know what was originally intended—so regulatory clarification is clearly needed. Nonetheless, regulatory refinement also has a downside. It is possible to "clarify" a law's regulations in a politicized rather than a procedurally helpful manner. Non-prescriptive statutory language, therefore, allows for both positive and negative consequences of subsequent regulatory explication.

If, however, those who craft a federal law spell out—in the law itself—the complete details of whatever is being required by the law, then there is less chance for subsequent misinterpretation and less need for subsequent clarification. But there is also less room for the ongoing adjustments that can make a law function more effectively. Congressional architects of assessment-related laws may be brimming with good intentions, but sometimes they make mistakes in framing a statute's particulars—especially when the particulars involve technical considerations such as those related to educational assessment. A prominent reason that prescriptive statutory language is particularly problematic in education-related legislation is because any federal law must allow for variations among 53 entities (50 states, the District of Columbia, Puerto Rico, and the Bureau of Indian Affairs). Creating statutory language so it is suitably circumscribing, yet sufficiently flexible, constitutes a nontrivial challenge to legislators.

A Three-Part Analysis

The bulk of this analysis will be divided into three sections. To preview briefly what's to come, a short description of each of those three sections will now be provided.

First, attention will be given to federal statutes that have had a meaningful impact on educational assessments in the U.S. during the last 50 years. Because these assessment-

related federal statutes could be addressed in enormous detail, and because there are space limitations associated with this analysis, it is my intention to provide only a very broad-brush recounting of the major assessment-related legislation of the past half-century, and the chief impact of this legislation.

Because I was asked to lay out options regarding how the federal government might implement a reconceptualized role in relation to the nation's educational assessments, in the second section of the paper I will attempt to describe what seem to be the most prominent control options available for influencing the nation's most significant educational tests. The resultant range of federal control options I will set forth refers exclusively to accountability tests such as those administered annually by states in order to comply with NCLB, or to satisfy the provisions of some similar federal legislation in the future.

Finally, whatever level of control option the federal government adopts regarding educational assessments, those options rumble into reality as soon as the actual tests must be developed. At this point, test talk is translated into the kind of test items we must use so we can arrive at inferences about what students know and can do. Accordingly, the third section of the paper will focus on three design dimensions that should be considered whenever accountability tests are to be built. Many of the accountability assessments now employed in America, without serious forethought by their builders, have taken on—sometimes by default—important functions related to our public schools' curriculum, instruction, or accountability. The third section of the paper will try to explicate the reasons that three dominant design dimensions should be given careful attention before the nature of any high-stakes educational assessment is determined.

A final, brief section of the paper urges the use of these two constructs, that is, the range of federal control options and the three design dimensions, as an organizing framework whenever serious discussions take place regarding a rethought role for the federal government regarding educational assessment. I know all too well how truly complicated it is to build educational assessments or to influence the nature and use of those assessments via federal or state statutes. But sometimes, in an attempt to deal with all of the complexities inherent in such endeavors, we lose sight of the most significant considerations in such an enterprise. To me, the two variables that make the most difference in this instance are (1) how much control over accountability tests the federal government should have and (2) what federally influenced tests should attempt to accomplish.

A BRIEF TOUR OF ASSESSMENT-RELEVANT FEDERAL STATUTES

Because decisions about if or how our federal government ought to influence America's educational assessments should surely take into consideration what has gone on legislatively in years past, it will be instructive to review, albeit briefly, the most salient federal statutes bearing on U.S. educational assessment. In turn, then, a once-over-lightly treatment follows of the National Defense Education Act, the Elementary and Secondary Education Act, the establishment of the National Assessment of Educational Progress, and the Education for All Handicapped Children Act.

Whereas it may be true that "what's past is prologue," past events are far from perfect predictors. A particular federal statute that could have a positive impact on the nation's schools if administered deftly by insightful federal personnel might prove disastrous when implemented ineptly by less talented federal officials. Federal laws are enacted at a particular time in a particular social context, and administered by particular people. As we

look back at previous federal legislation related to educational assessment, we may encounter some statutes that appear to have been less successful than their supporters had anticipated. This does not signify that all similar legislation is destined to stumble. We should surely try to learn from history, but not all legislative history is laden with exportable lessons.

The National Defense Education Act (NDEA)

A threatening Cold War served as backdrop for the enactment of the 1958 National Defense Education Act. It is generally conceded that the Soviet Union's launching of Sputnik finally confirmed an increasingly widespread belief that America was technologically falling behind the Soviets. To cope with this perception, our public schools were viewed as both the problem and the remedy (Anderson, 2005). Educational assessment, however, played only a small role in this legislation. Title V of NDEA provided grant monies to states for "guidance, counseling, *testing* [emphasis added] and identification of able students" (Spring, 2008 p. 402). In addition, modest funding was provided for states to develop data-gathering and reporting systems (Spring, 2008). Otherwise, NDEA appears to have had little impact on U.S. educational assessment.

.

The Elementary and Secondary Education Act (ESEA)

In marked contrast, the Elementary and Secondary Education Act has dramatically altered the federal government's role in U.S. public education, especially with regard to the assessment of student achievement (Anderson, 2007). ESEA has been a significant part of America's educational landscape since its enactment in 1965 (Anderson, 2005; Jennings, 2001), and has been reauthorized eight times[*] (Anderson, 2005). ESEA has provided billions of dollars for the nation's public schools.

Congress first authorized ESEA as a part of President Lyndon Johnson's War on Poverty (Jennings, 2001). The law's purpose was to "provide financial resources to schools to enhance the learning experiences of underprivileged children" (Thomas & Brady, 2005, p. 51). At the beginning, and until its 1994 reauthorization as the Improving America's Schools Act (IASA), educational assessment in ESEA was largely limited to a form of fiscal monitoring in order to ensure that federal dollars were effectively serving the intended populations (Anderson, 2007; Jennings, 2001; Riddle, 2001).

The original 1965 law contained several provisions, or "titles," but Title I was by far the most recognized and influential. Title I was the provision supplying categorical funding for low socioeconomic status (SES) school-aged children. The law did make an attempt at evaluation (Merkel-Keller, 1986), noting that state education agencies (SEAs) and local education agencies (LEAs) could only receive Title I funds if there were "procedures for evaluating the effectiveness of the programs . . . in meeting the special education needs of educationally deprived children" and that the "evaluations will include effective measurements of educational achievement in basic skills" (Title I, Section 172 cited in Merkel-Keller, 1986, p. 8).

Section 172 of ESEA also included a reporting requirement whereby SEAs were to "make periodic reports . . . evaluating the effectiveness of the program" (Title I, Section 172, cited in Merkel-Keller, 1986, p. 8) to the Secretary of Education assuring that Title I students were making progress. Difficulties arose in that LEAs and SEAs had no consistent way to assess

[*] ESEA was reauthorized in 1968, 1972, 1974, 1978, 1982, 1988, 1994, and 2002.

students, collect data, or report the findings. Nor were there adverse consequences for failing to do so, thus rendering the provision ineffectual. However, this evaluation provision of ESEA could be regarded as a significant step toward today's federal assessment, evaluation, and accountability requirements.

The 1974 reauthorization of ESEA saw assessment gain ground. Concerns over the accounting and monitoring of the use of funds receded, and Congress turned its attention to the impact of ESEA on student achievement (Borman, 2000; Borman & D'Agostino, 2001; Borman & D'Agostino, 1996; Merkel-Keller, 1986). The 1974 amendments included funding provisions for the Secretary of Education to develop an independent and systematic evaluation and reporting system to determine the educational effectiveness of ESEA with "objective criteria . . . utilized in the evaluation of all programs . . . producing data which are comparable on a statewide and nationwide basis" (Education Amendments of 1974, 88 Stat. 484, 500 as cited in Borman & D'Agostino, 2001, p. 41). Between 1974 and 1978, the Title I Evaluation and Reporting System (TIERS) was developed. TIERS implementation began in 1979 after the 1978 reauthorization of ESEA.

TIERS consisted of a three-level process for evaluating and reporting (Borman, 2000; Borman & D'Agostino, 2001); Merkel-Keller, 1986; Rutherford & Hoffman, 1981). At the first level, TIER I, the local education agency collected student achievement data and reported this information to the state education agency. In TIER 2, the SEA aggregated the LEA data and reported these results to the commissioner of education. For TIER 3, the U.S. Office of Education aggregated the nationwide data and reported on the effectiveness of Title I programs in meeting the needs of the identified disadvantaged population.

The implementation of TIERS generated major methodological issues. For example, there were concerns about the accuracy of the cumulative national data. Because SEAs and LEAs retained control over the selection of achievement tests, there were a number of different tests employed throughout the states. To illustrate, states used such assessments as the *Stanford Achievement Tests*, the *Comprehensive Tests of Basic Skills*, and the *Iowa Tests of Basic Skills*. Multiple tests, of course, led to questions regarding the legitimacy of between-state comparisons. Concerns over the administration of the tests also began to surface. Were guidelines for administration the same across all schools in an LEA as well as throughout a state? There were also questions regarding the timing of test administrations, because SEAs and LEAs often tested students at different times of the year. Although federal officials attempted to address these methodological issues in various ways, many of those technical solutions proved to be less than satisfying.

In the 1988 Hawkins-Stafford Act reauthorization of Title I, accountability requirements were further strengthened. The act required states to "define the levels of academic achievement disadvantaged students *should* [emphasis added] attain in schools receiving Title I funds" (Jennings, 2001; McDonnell, 2005). The 1988 version of ESEA required SEAs and LEAs not only to develop plans for increasing student performance, but also to test and report student achievement of the Title I population *annually* (Riddle, 1989; Thomas & Brady, 2005). While many states had by this time implemented annual assessments of their students, this was the first time that yearly administrations were required (Thomas & Brady, 2005). Moreover, there was a new evaluative provision for those schools that elected to use their Title I funds on a schoolwide basis. This provision required that after a three-year implementation, such schools must demonstrate that the achievement of their disadvantaged children was higher than either the average of children participating in the LEA as a whole, or [was higher than]

the average for disadvantaged children in that school over the three years preceding the schoolwide plan's implementation (Riddle, 1989, p. 11).

IASA, in a legislative change of considerable import, shifted its evaluative focus toward a more clear delineation of educational expectations and a more comprehensive effort to assess students' attainment of those expectations (The Commission on No Child Left Behind, 2007; DeBray, 2005; Hanushek & Raymond, 2005; Jennings, 2001). To qualify for IASA Title I funds, "states were required to adopt and implement curriculum content standards, pupil performance standards, and assessments linked to these" (Riddle, 2001, p. 2). In addition to the development of standards and assessments of students, IASA required that *all* students be assessed, especially second language learners and students with disabilities. IASA, a particularly significant reauthorization of ESEA, also required that student achievement data be disaggregated for statute-identified subpopulations (Fast & Erpenbach, 2004).

A major modification in IASA was its effort to move states away from reliance on norm-referenced measurement toward the use of a criterion-referenced assessment approach. Riddle (2008) notes this significance, stating these "two types of tests vary *primarily* [emphasis in the original] regarding how test results are analyzed, but also typically differ to some degree with respect to such characteristics as the range of questions included" (p. 3). Therefore, because IASA required that assessments be tied to standards and to determining the achievement of those standards—rather than comparing students with one another—many state officials found that they needed to move toward a starkly different assessment strategy for their state's accountability tests. As states attempted to accomplish this change, Redfield & Sheinker (2004) note that "approaches range from adding items to available norm-referenced tests (NRTs) to align them with state academic content standards and grade-level expectations (GLEs) to creating new standards-based tests (p. 8)."

Passage of the No Child Left Behind Act, the 2002 reauthorization of ESEA, dramatically magnified the assessment requirements of IASA by adding significant stipulations regarding assessment (DeBray, 2005; Riddle, 2008; Shaul & Ganson, 2005) and by installing a series of potent sanctions for schools and districts whose students failed to sufficiently improve their scores on a state's annual accountability tests.

Rather than being obliged to administer assessments in a grade *range* (e.g., grades 3-5) as had been called for in IASA, states were required by NCLB to develop and/or adopt assessments for *each* of grades 3-8 and once in grades 9-12 in the areas of reading/language arts and mathematics. NCLB also required that science be added for grade ranges 3-5, 6-9, and once in high school (Erpenbach, Forte-Fast, & Potts, 2003; Mills, 2008; Riddle, 2008). In a sense, NCLB took many of the requirements of IASA and put sharper teeth into them. The disaggregation of subgroup test scores in NCLB, for example, could lead to far more serious evaluative consequences than had been the case for IASA's subgroup disaggregation. Many observers regard NCLB's more stringent demands for subgroup disaggregation to be a particularly significant contribution of this most recent reauthorization of ESEA.

In what may turn out to be one of its more significant changes, NCLB required states and those local districts selected in the sampling plan to participate in the National Assessment of Educational Progress (NAEP) in the 4^{th} and 8^{th} grade in mathematics and reading (Redfield & Sheinker, 2004; Shaul and Ganson, 2005; Riddle, 2008). This was a major change, because until NCLB, participation in NAEP had been voluntary (Riddle, 2008; Shaul

& Ganson, 2005). With the availability of state-by-state NAEP scores, it was apparent that NCLB-required performances of students from different states would most certainly be compared. NAEP would, in the eyes of many policymakers, become the *de facto* yardstick by which the success of a given state's public schools would be determined.

The National Assessment of Educational Progress (NAEP)

Authorized and funded by Congress when initially implemented in 1969, the National Assessment of Educational Progress (often referred to as "the Nation's Report Card") provides information to lawmakers, policymakers, and the general public about U.S. students' academic achievement (Bowers, 1991; Hombo, 2003; Riddle, 1998). Since its inception, NAEP has provided a "snapshot" of American students' achievement. In its early years, NAEP was a particularly innovative assessment program, annually assessing diverse content areas with a significant proportion of its tests consisting of performance assessments. Because of NCLB requirements, since 2003 NAEP has provided results in reading and mathematics at grades 4 and 8 every two years for states and the nation. Results are not available by school or by student, because (with the exception of several large school districts) the smallest NAEP reporting units are states.

NAEP is not one test, but rather a set of three tests that each serve a specific purpose. The NAEP trend test is given in reading and mathematics every four years to student samples at ages 9, 13, and 17 (http://nces.ed.gov/nationsreportcard). The trend test has provided long-term, comparable information about nationwide student achievement since its inception in 1969. Results are reported for the nation as a whole and for subpopulations such as gender and ethnicity. The main test, given every two years in grades 4, 8, and 12, provides measurements of student achievement in the areas of reading and mathematics. As with the trend test, results are reported for the nation as a whole and for subgroups (http://nces.ed.gov/nationsreportcard). State tests, the third type of NAEP test, were added in 1990 based on a perceived need for comparable data across states (Hombo, 2003). These assessments are administered in the opposite years from the main tests in the areas of reading, mathematics, and writing.

NAEP appears to have had an impact on the nature of America's assessments as a whole, and is thought by some to have increased the public's confidence in testing (Bowers, 1991). For example, NAEP results have been used to evaluate the success of programs such as NCLB (Fuller, Wright, Gesicki, & Kang, 2007; NCES, 2006; Center on Education Policy, 2008) and "to study the effect of accountability on student performance (Herman, 2007, p. 11)." NAEP's assessment procedures have also served as models for numerous state testing programs (Campbell, Hombo, & Mazzeo,1999).

Not all observers of NAEP are fans. For example, in a review of a 2004 anthology about NAEP (*The Nation's Report Card: Evolution and Perspectives),* Stake registers strong doubts about the ultimate benefits of this oft-evaluated program:

> With strong support from the measurements community, the main characters in this book about NAEP created NAEP in their own image. They wanted it to be the best that they could be. To be pure assessment, they disdained curriculum experts and philosophers. But they failed to demand validation of the assessment's core policy. At first, the core policy was tracing performance over time, but gradually the core policy became test-based accountability. (Stake, 2007, p. 18)

At the very least, the methodological measurement refinements emerging from NAEP have surely had a considerable influence on the way many of America's educational assessments are currently created. Given the increasing importance attributed to a state's NAEP scores by educational policymakers, we may well see this assessment program, almost 40 years old now, attain a position of ever-increasing prominence in the way states choose to assess their students.

To illustrate, if NAEP results become regarded as the most credible way of evaluating a state's schools, then is it not likely that at least some state officials will urge their state's assessment personnel to "measure what NAEP measures?" Even though, as long as NAEP scores are not reported at the school level, NAEP will remain a relatively low-stakes assessment, its curricular impact at the state level could thus be considerable.

The Education for All Handicapped Children Act (PL 94-142) and the Individuals with Disabilities Act (IDEA)

The Education for All Handicapped Children Act (PL 94-142) of 1975 and its reauthorized version, the Individuals with Disabilities Act of 1990[†] (IDEA) has had a substantial impact on the way many thousands of American children have been educated (Itkonen, 2007; Koegh, 2007); Thurlow, Lazarus, Thompson & Morse, (2005). Until the enactment of PL 94-142, children with disabilities had limited or no access to public education (Lehr & Thurlow, 2003; Zettel & Ballard, 1979). PL 94-142, however, mandated that children with disabilities be provided a "free and appropriate public education" (FAPE) and in the least restrictive environment, which usually meant the regular classroom (McLaughlin & Nagle, 2004). PL 94-142, as had no previous federal legislation, ensured equal access to public education for all children with disabilities (Gaddy, McNulty, & Waters, 2002; McLaughlin & Nagel, 2004).

In the early years of the law, its assessment and evaluation provisions focused on compliance and monitoring (Anderson, 2005; DeStefano, 1992; Gaddy, McNulty, & Waters, 2002; Lehr & Thurlow, 2003; McLaughlin & Nagle, 2004). However, when PL 94-142 was reauthorized as IDEA in 1997 and again in 2004, its measurement focus changed from assessing how the program was being implemented to instead determining the academic achievement of the populations served (Lehr & Thurlow, 2003; McLaughlin & Nagle, 2004). Put simply, the assessment emphasis of this legislation shifted from inputs to outputs. IDEA also mandated that state assessments for children with disabilities be tied to the same state standards that had been adopted for other children in that state.

With the shift to assessing academic achievement, IDEA mandated that students with disabilities be included in the same general assessment as their nondisabled peers. No longer could disabled students be excluded from the general testing programs on the basis of their disabilities (McLaughlin & Nagle, 2004; Thurlow, Lazarus, Thompson & Morse, 2005). Furthermore, test results for students with disabilities were to be reported as a part of the total-population aggregate and again as a separate subgroup (Riddle, 2008; Thurlow, Lazarus, Thompson & Morse, 2005).

Perhaps the greatest impact of IDEA on educational assessment revolved around the measurement of students who, due to significant cognitive disabilities, were unable to participate in the general testing. Because all students were to be tested, IDEA required that

[†] IDEA was amended and/or reauthorized in 1991, 1994, 1997, and 2004.

an alternate assessment for this population must be developed (Lehr & Thurlow, 2003; Thurlow, Lazarus, Thompson & Morse, 2005), thus requiring all states to create such assessments. However, the 1997 IDEA did not specify the content or form of these assessments (Hager & Slocum, 2002), nor was this topic addressed in the 2004 reauthorized statute (DOE, Office of Special Education Programs, 2007). As a result, states have developed their own alternate assessments, and while there is no mandate regarding content, Lehr and Thurlow (2003) report that there has been a shift from "functional skills to student achievement of state standards."

The form of the alternate assessments varies. Different states have developed different types of alternate assessments for their students with significant cognitive impairments. These alternate assessments include portfolios, checklists, rating scales, and sometimes even rely on analyses of students' Individual Educational Plans (IEPs) as the alternate assessment (Hager & Slocum, 2002). In earlier years, the assessment of children with disabilities was regarded by most general-education teachers as a task only of concern to their special-education colleagues, but this is no longer the case. Rather, because many children who have special needs now are included in regular classrooms, today's teachers are confronted far more frequently with the need to arrive at appropriate assessment approaches for children with disabilities. Clearly, these changes in the way special-needs children are currently assessed have been brought about by a series of landmark federal laws regarding how to educate and how to *assess* those children.

Looking back, then, at this brief overview of a half-century's worth of federal legislation bearing on educational assessment, we have seen an ever-increasing impact of these federal laws on real-world assessment practices. Early versions of ESEA, for example, found educational testing used chiefly for monitoring whether federal funds for statute-specified populations were being effectively used for those populations. But in the early nineties, an important shift in purpose was seen when ESEA assessment results were to be used in making sure that all students, not just the disadvantaged students identified in Title I of the law, were achieving the curricular aspirations that had been set out for them. Moreover, the curricular aims for children with disabilities—albeit modified to some extent—increasingly became identical to the curricular aims sought for *all* students. This shift in orientation took place gingerly in IASA and in early versions of IDEA, but arrived with considerably more clout in the most recent renditions of both ESEA and IDEA.

At this moment, there seem to be few challenges to the belief that educational test results should be used to ascertain how well the nation's students have been taught. Whether the federal government *should* play a prominent role in determining the nature and uses of U.S. accountability tests is, of course, a pivotal consideration in any rethinking of an appropriate federal role related to educational assessment.

A RANGE OF FEDERAL CONTROL OPTIONS FOR ACCOUNTABILITY TESTS

In recognition of the considerable impact that accountability tests can have on public schooling, American policymakers—especially during the past few decades—have begun to deliberate how much sway the federal government should have in determining the nature and use of the nation's most significant accountability assessments. This issue, interestingly, is a relatively recent one to surface. In the absence of a Constitutional stipulation to the contrary, U.S. public education has traditionally been seen as the purview of state and local authorities—not that of the federal government.

If we were somehow able to time-travel back for 50 or so years, we'd find that any serious suggestions of *federally* controlled education would be instantly rebuffed by almost all American citizens, including almost all educators. Education was so universally thought to be the proper realm of nonfederal authorities that adherence to "local control of education" represented one of our nation's almost sacrosanct beliefs. But of course, times change.

During the past 50 years, antipathy toward federal control of education—tradition notwithstanding—has been diminishing. Growing numbers of Americans have concluded that the nation's long-time state control of schools has led to an almost anarchic situation in which what we expect of children in one state's public schools is often dramatically different from what we expect of children in another state's public schools. Moreover, it is all too apparent that substantial differences in instructional quality exist among our states. Children who live in certain states receive a less effective education than do children in other states. Taken together, these state-to-state differences in expectations and in instructional quality have led increasing numbers of policymakers to a call for the federal government—not states—to determine how the nation's students are tested. A serious federal influence on America's public schools that barely a generation ago was regarded as anathema is now seen by many stakeholders as a potentially viable alternative.

So, unlike in years past, anyone who is now seriously trying to rethink a suitable federal role in educational testing is able to consider a set of options that previously would have been almost unthinkable. A number of citizens and educators wish to see the federal government exercise a much greater influence on what gets tested, hence gets taught, in our schools. Indeed, the dramatic impact of NCLB assessment requirements on what goes on in our schools has shown us all, in only a half-dozen years, how a set of federal preferences regarding educational assessments can make a marked difference in what's tested and, therefore, what's taught in the U.S. public schools.

Interestingly, discussions of the impending reauthorization of NCLB (the current incarnation of 1965's ESEA) have failed to engender widespread agreement regarding what the assessment provisions of this soon-to-be reauthorized statute should look like. As a consequence, uncertainty now exists regarding what kind of federal influence over educational assessments should be incorporated in any reauthorized version of NCLB. So, given what appears to be at least some erosion of opposition to strong federal influence on public schooling, coupled with the current diversity of views about what federal influence on assessment should be built into in a reauthorized NCLB, the timing for a serious rethinking of the federal role in U.S. educational assessment could not be better. While we may not be dealing with a true *tabula rasa,* our educational assessment slate seems sufficiently clean so that, if we choose to do so, we can think relatively unfettered thoughts about what kind of federal influence on high-stakes educational tests would best benefit the nation's students.

Absent any pre-existing constraints about what levels of federal influence on educational assessment are worth considering, it is theoretically possible for the federal government's control of educational assessment to range from *total* control over the nation's educational accountability tests all the way to the opposite extreme, where the federal government has *zero* control of those tests. In an attempt to illuminate the scope of options that policymakers ought to consider as they rethink an appropriate level of federal involvement in the nation's accountability testing, I will briefly describe a series of five alternate *control options* that might represent a suitable role for federal involvement in educational assessment. The resulting continuum of federal influence on educational assessment, ranging from none to total, can be regarded as a

range of federal control options. Let's get underway, then, with a description of Control Option 1, an alternative that is, in essence, a *no-control* option.

Control Option 1: Absence of Federal Control

At one end of a range of federal control options is the total absence of federal control, that is, no control whatsoever over the accountability tests that have such a powerful impact on what goes on in our schools. We can designate this alternative as Control Option 1. Lest it be thought that such a position is more theoretical than real, simply consult the April 2008 *Phi Delta Kappan,* where Phillip Schlechty, a seasoned analyst of American education, calls for an abandonment of any federal influence on educational accountability operations. While not opposed to federal aid, Schlechty wants federal dollars disbursed in the form of block grants rather than as categorical funds, thereby allowing local officials to decide how to evaluate their own instructional effectiveness. As Schlechty points out:

> The type of accountability plan local communities adopt should be left up to the local school boards, with the stipulation that each district would conduct an annual survey of all citizens regarding their level of satisfaction with the performance of the schools, as well as with the quantity and quality of the information regarding school performance (Schlechty, 2008, p. 557).

Those who yearn for an end to federal authority over educational assessments are apt to agree with Schlechty when he rejects a key assumption of federally imposed accountability "that standards imposed from outside a system can inspire excellence within a system even as they ensure that performance does not fall below some minimum level" (Schlechty, 2008, p. 556). Proponents of a control option in which federal influences are nonexistent do not, as we see from Schlechty's comments, reject all forms of educational accountability. Rather, the advocates of Control Option 1 would prefer that any such accountability take place at the local, or possibly the state level, with or without federal funding.

Each of the five control options to be considered here can be implemented in diverse ways. For example, because the emphasis of Control Option 1 is on the complete absence of federal *control*, a federal approach to educational assessment stressing information and professional development should still be regarded as a legitimate variant of Control Option 1. A federal law, for instance, might provide for substantial federally supported professional development focused on the more effective use of formative assessment. State or district educators could participate actively in such *gratis* professional development endeavors advocating the expanded use of formative assessment, but after such participation might decide, without penalty, not to promote what federal authorities had been recommending. The proffering of assessment-relevant information and professional development is not the same as *control.*

If Control Option 1 were to be adopted—an approach in which there was a total absence of federal influences on accountability testing—it might be necessary either to eliminate NAEP altogether or to dramatically restructure it so that, as in earlier days, no state-by-state comparisons were possible. As long as NAEP exists in its current form, whereby its state-by-state results are seen by many as the "real" indicators of how well a state's students are progressing, NAEP is certain to influence the way state or local governments assess their students' achievement. State education officials, quite predictably, will be reluctant to have their state's test results be meaningfully out of line with results of "the nation's report card." As a result, we may see some of these "outlier" states trying to create their own NAEP-like accountability tests and adopt NAEP-like passing standards so their state's test performances

will be more consonant with those of NAEP. Adoption of Control Option 1 might necessitate a dismantling of NAEP, or at least a "softening" of key provisions in that federal testing program so that its impact on state accountability programs would be minimized.

It would always be possible, of course, to enact federal laws that eliminated all federal influence on state accountability tests, yet still retain NAEP as-is in its current form, whereby state-by-state comparisons are not only possible, but often form the grist for a state's education leaders to either crow or cower. This sort of hybrid approach may certainly be selected, but it would not be an example of Control Option 1, in which there must be a *complete* absence of federal influence on a state's high-stakes tests.

Hybrid approaches to federal control of accountability tests clearly illustrate that the five alternatives represented in this paper's range of control options need not be adopted intact or in isolation. The five control options presented here are intended to help clarify several relatively distinct alternatives that might, with or without modification, be embodied in subsequent federal legislation. Combinations of two or more control options described here, or even meaningfully modified versions of any of those five control options, are certainly possible. For instance, federal lawmakers might decide to adopt one control option for certain assessment functions, yet adopt another control option for other assessment functions. To repeat, the purpose of the range of control options set forth in this paper is solely to make more vivid the choices educational policymakers face regarding the degree to which federal forces should shape the nation's most significant accountability assessments.

Control Option 2: Federally Provided Optional Use Tests

Moving toward greater federal impact on U.S. accountability assessments, Control Option 2 calls for the federal government to generate a wide array of excellent assessment instruments, then make those tests available to states (and/or districts), but only for voluntary use, that is, only if state (or district) officials choose to use such tests. The positive feature of this control option is that substantial federal resources could be brought to bear on the nontrivial task of creating first-rate accountability tests. As matters currently stand, despite zealous efforts to create excellent assessments for a state's accountability program, some state-level accountability tests are far less wonderful than their creators had hoped. If the federal government, perhaps drawing on the expertise of several federally supported laboratories and R&D centers, could create a crackerjack collection of accountability assessment tools to measure students' mastery of the most commonly sought curricular aims, then these tools could be adopted and/or adapted by state or district assessment officials.

Federal authorities would be influencing the accountability assessments used in the U.S. by supplying ready-to-use, "turn-key" exemplars representing outstanding assessment instruments. The better the quality of the federally created assessment devices, the more likely it seems that those assessments might be adopted by state officials. In essence, Control Option 2 features the provision of a collection of optional-use measurement tools for the nation's educators. Because different states have different curricular emphases, this control option would work best if a substantial variety of assessment devices were created, thus making it more likely for states to find assessments in line with their state's particular curricular emphases. In the same vein, to provide for more appropriate alignment with different states' curricular emphases, at least some of these federally fostered assessments could be created in modular form, thereby allowing for states to engage in their own mix-and-match use of selected assessment modules.

Consideration of Control Option 2 does not suggest that all current state-developed accountability tests are inadequate. A number of states have expended considerable energy in providing their state's educators with first-rate assessments. Yet, given the limited financial and/or personnel resources of many states, the possibility of adopting at least some assessments constructed under federal funding may be attractive to authorities in certain states.

The generation of high-quality accountability tests is far from fool's play. By marshalling the considerable fiscal resources of the federal government, it should be possible to fashion truly stellar versions of accountability tests that would have a positive impact on a state's instruction, curriculum, and/or accountability activities. Perhaps for each *accountability* test created under federal auspices a set of related assessment instruments could be developed that would be intended for use as part of the *formative* assessment process linked to the curricular aims being assessed in the accountability test. If Control Option 2 were to be adopted, with its main thrust being to generate a galaxy of top-drawer assessment tools, then this control option might command more than mere dollars. It might be possible, at least for a period of several start-up years, to attract some of the nation's top assessment experts who would, by taking part in this pivotal federal project, be able to play a salient part in improving our nation's education. Such an opportunity might entice many of our finest assessment specialists to take part in this sort of test-building enterprise. Even assessment specialists can be lured by the right kind of crusade.

Once a state had chosen one or more of these federally crafted accountability assessments, of course, there would be the practical matter of replenishing those instruments so that future test forms would contain sufficient numbers of new items. This replenishment task could be addressed in a variety of ways, one of which might be for a state—perhaps with some level of financial support from the federal government—to take on such a replenishing task by itself. Clearly, once the specifications for a particular accountability test at a particular grade level have been worked out, it would be far less expensive for an independent agency to subsequently provide replacement items than if those specifications had not been created.

It would seem likely that, were Control Option 2 to be adopted, the administration, scoring, and reporting of these accountability tests would be carried out by independent, nongovernmental assessment firms, much as is currently the case in all but a few states. For item replenishment of such tests, those external firms would surely be able to follow the item-generation and test-assembly specifications created by federal projects. New test forms, replete with sufficient form-equating linked items, could be developed under state auspices with or without federal financial subvention.

Because the adoption of these federally created assessments would be totally voluntary for states, this would also present an opportunity for federal personnel to prepare a variety of concomitant materials and procedures that could, *if voluntarily chosen by a state's officials,* support the intended mission of the particular assessment instruments selected. Suppose, for instance, that a set of optional-use federal accountability tests in mathematics had been developed for use with a state's elementary-school students. The dual mission of these tests might be to (1) provide credible per-school accountability evidence so that a school's instructional effectiveness could be properly evaluated, and (2) improve the quality of mathematics instruction taking place in a state's elementary schools. Given such a scenario, it would surely be possible for federal personnel to generate a number of materials and implementation procedures in support of both of those aims. As always, whether state or local educators chose to employ any such materials or procedures would be up to those state or local educators.

The thrust of this second control option is obviously to build better mousetraps so that those in need of measurement mice-trapping will have an array of superior accountability assessments to employ. The federal government would not force anyone to use its carefully constructed instruments and their ancillary support materials, but to the degree that those tests and supportive materials were regarded with favor by local and state educators, improved accountability testing is almost certain to take place in many locales. If states are supplied with some level of federal financial support as an incentive for adopting such tests, of course, this would meaningfully increase the acceptance of Control Option 2.

Control Option 3: Statute-Required State Tests with Light Federal Control

As we move toward more federal influence along our "none-to-lots" range of control options, we begin to see versions of federal control that actually resemble the way in which recent federal statutes have attempted to exercise influence over the nation's accountability tests. Control Option 3 calls for the enactment of federal laws requiring states to use accountability tests, but those tests can be of a state's own choosing. However, by linking federal funds to the stipulated nature of the tests, and to the way a state must use these tests, federal personnel can clearly have an influence on these statute-stipulated state accountability tests. In Control Option 3, however, this federal control would be only modest. To some degree, this particular control option is similar to the way that IASA, the 1994 reauthorization of 1965's ESEA that immediately preceded NCLB, dealt with state accountability assessments.

In IASA, states were required to select and administer their own accountability assessments at three grade ranges, but the nature of those tests was not controlled rigidly by federal authorities. Indeed, even though state assessment personnel had been adjured by federal officials to build and use their state accountability assessments in certain ways, and federal personnel had supplied directives regarding how the IASA accountability assessments ought to be both built and administered, at the time IASA was soon to be replaced by NCLB, only about one-third of the states had secured federal approval for their IASA accountability programs (Fast and Erpenbach, 2004).

Prior to IASA, a number of states actually had no state assessment programs, and though prodded by IASA's "light" federal influence, were slow to develop one. We see, therefore, that although there were federal recommendations regarding the way state accountability tests ought to be built and used, only modest federal pressure was employed to get states to play their accountability games—and secure their federal funds—according to IASA ground rules. This is an example of how the federal government can require tests to be built by states, yet allow considerable latitude regarding how those tests are constructed and how they are to be used.

In this particular control option, of course, the degree to which federal personnel can tighten or loosen any statutory constraints on (1) the nature of a state's accountability tests or (2) how those tests are to be used would depend on the specific language in the law dealing with these tests. As noted earlier, such statutes vary considerably in how much interpretive license they provide for federal implementation.

In essence, Control Option 3 boils down to a federal requirement that states create and/or select and use accountability tests, but Control Option 3 also embodies a concomitant federal commitment to influence only mildly what the tests are like and how those tests are to be used. This approach, of course, allows for greater flexibility on the part of state educational officials

and their staffs. To some, this sort of flexibility is seen as a clear plus. In contrast, others regard such flexibility as Control Option 3's major deficit.

Control Option 4: Statute-Required State Tests with Tight Federal Control

The next alternative, Control Option 4, bears a striking similarity to the way NCLB has been administered by federal personnel since its enactment. The heart of this fourth control option is that a federal statute establishes the requirement for states to select their own accountability tests (either custom-built or chosen from available commercial assessment alternatives), but then the regulations and guidances emanating from this federal statute are employed to exercise a strong influence on how a state's officials attempt to comply with the statute. In NCLB, we find a federal law calling for more than twice the number of required accountability tests previously demanded from states. Moreover, the uses of the test results are tightly prescribed by regulations and in subsequent guidance. To illustrate, the peer-review process established by the U.S. Department of Education promulgated a series of explicit test-appraisal criteria that in many states triggered significant alterations in the way a state was building and/or using its accountability tests.

If one were to render a gestalt overview of what the federal assessment demands on states were based on IASA versus what those assessment demands on states were based on NCLB, it would be apparent that federal assessment demands were ratcheted up dramatically by NCLB. In both of those ESEA reauthorizations there were federal requirements for annual state testing, and in NCLB the numbers of those required tests more than doubled beyond what was called for in IASA. But in NCLB we saw a major shift in the degree to which federal officials—based on the NCLB statute itself—set out to make sure that state assessment personnel were selecting the right accountability tests and were using those tests and their results in very specific, federally dictated ways.

When the federal government can withhold considerable dollars from a state education agency if that state deviates from the kind of testing that federal personnel believe is stipulated by a federal law, and if those federal officials wish to control the nature of state accountability tests and how they are used, then federal officials are in an excellent position to implement some variant of Control Option 4. Since the enactment of NCLB in early 2002, this is precisely what authorities in the U.S. Department of Education have been doing, making certain that state accountability tests cleave to the NCLB statute, its regulations, and its guidance. We have, accordingly, a recent and vivid illustration of how Control Option 4 would function.

Why, some might ask, would federal personnel seek to control with so much energy the NCLB-dictated accountability testing enterprise? After all, during IASA there had been at least some constraints on state accountability tests and their use. Why, then, given whatever implementation latitude was present in NCLB and its regulations, did federal authorities tighten the assessment requirements so stringently?

One possible explanation is that, under IASA and the many previous incarnations of ESEA, numerous state and local educators adopted a series of "gaming" ploys to circumvent many of the law's constraints that might make those educators appear unsuccessful. For example, IASA anecdotes are legion about districts' sending potentially low-scoring students on "field trips" during the days when IASA accountability tests were due to be administered. Accordingly, in NCLB we find a new requirement that at least 95 percent of eligible students must complete all of the NCLB accountability assessments. Many similar tightened-oversight provisions are found in NCLB and its regulations. What comes through from a reading of the statute and its

regulations is quite clear: This was to be a law that educators could not easily escape. This was a law to be obeyed—or those educators who disobeyed would suffer serious consequences. As such, NCLB is a poster child exemplar of Control Option 4.

Obviously, the particular constraints in a federal accountability statute will play a prominent role in any rendition of this kind of control option. NCLB, for instance, called for serious scrutiny of the academic content standards that were to serve as the focus of the law's accountability approach. Moreover, an attempt was made to oblige state officials to come up with solid evidence that those content standards, if suitably selected, would be properly measured. There are, indeed, myriad particulars to be considered in carrying out Control Option 4. But the essence of this approach remains unmistakable. In this control option, one or more federal laws require state use of accountability tests, and exacting federal rules oblige states to play the assessment game using a strict federal playbook.

Control Option 5: Complete Federal Control

In contrast to one end of the range of control options in which there is a total absence of federal control over the nation's accountability tests, we now come to Control Option 5, in which federal control of those tests is *complete*. Were this control option to be enacted by federal law, then essentially all governmentally required accountability tests would be the responsibility of federal authorities. In practical terms, this would mean that a federal agency, most likely the U.S. Department of Education, would be in charge of building, administering, scoring, and reporting results of those tests. A federal statute authorizing this kind of arrangement could, of course, allow for some state-level collaboration or the assistance of private assessment organizations in order to carry out such a huge assessment operation. But the final decisions regarding all aspects of such an enormous assessment operation would be the responsibility of federal personnel.

Illustrative of the position of those who favor Control Option 5 is a recent essay, "The Case for National Standards and Testing," in which Walt Gardner draws a parallel between today's situation in America's schools and Abraham Flexner's landmark evaluation of medical schools. In his 1910 report, Flexner pointed out that because the schools differed so widely in their curricula, methods of assessment, and graduation requirements, "it was impossible to know with any degree of certainty if students were being well educated" (Gardner, 2008, p. 26). Gardner opines that the current situation in our public schools is not unlike that of medical schools a century ago, before the introduction of qualifying exams increased curricular comparability in medical schools. Gardner contends that it is time for the adoption of national standards and for the use of national tests to see if students have mastered those national curricular aims. Faced with an American educational system some observers regard as chaotic, we can understand why Control Option 5 might be attracting larger numbers of proponents.

Any authorizing legislation embodying this kind of control option must deal with the kinds of statutory constraints discussed at the outset of this analysis. For example, a federal law intended to establish Control Option 5 would need to spell out precisely what tests are to be administered, at what grade levels, at what times of the school year, and so on. Or instead, that law might simply authorize the establishment of some sort of governing group that would be empowered to make such substantive and procedural decisions after seeking advice from concerned constituencies. The means of implementing a control option in which the federal government takes on total control of educational testing would obviously be complex and challenging, but—at least *conceptually*—need be no more complicated than are the problems

encountered when assessment personnel in each of the 50 states must build, administer, score, and report results for their own state-level accountability tests.

There are important conceptual similarities between state-level and national-level assessment operations. One of the first tasks that must be accomplished when building state-level accountability tests is to decide what is to be assessed. To do this, state officials typically look to the academic content standards their state has already adopted, the set of curricular aims describing the skills and knowledge the state's students are supposed to acquire. Because these curricular aims will supply the assessment direction for a state accountability test, great care is taken regarding their determination. Were there to be a set of national tests created anew, then similar—but far more intense—scrutiny would surely be given to the determination of a set of national curricular aims, that is, a set of national academic content standards. At almost every point in the creation and use of state accountability tests, a comparable operation (hopefully much more rigorous) would be required if national tests were to be developed. Given the number of players who would be taking part in any national push toward curricular consensus, it would seem that a lengthy period of curricular determination—perhaps several years—might be required.

One approach to the implementation of Control Option 5 would be the creation of brand-new accountability assessments, in the same way that most of our states were required by NCLB to build new tests for grade levels where no such tests existed. But it is also possible that a law establishing Control Option 5 would simply expand the scope of NAEP so that this well-established enterprise could take on the functions called for by any new assessment-engendering legislation.

Changes would most certainly be needed in NAEP were it to be chosen as the assessment vehicle to implement this sort of control option. For example, additional grade levels might be assessed and additional subjects might be measured. The frequency of NAEP administrations might also be increased. Then too, because NAEP currently employs a matrix-sampling administration design in which different students complete different test forms, and because parents have usually registered strong dissatisfaction when state tests have used this approach (because individual students' performances on different test forms can't be sensibly compared), a decision might be made to abandon matrix-sampling models for any sort of expanded, accountability-focused application of NAEP. Technical advances in online administration and scoring of tests, of course, could play a significant role in this sort of national assessment.

But whether this control option, an approach reflecting complete federal control of accountability assessments, would be implemented with or without reliance on NAEP, the thrust of Control Option 5 is unmistakable. If this strategy for carrying out U.S. accountability testing were adopted, there would be national curricular aims being measured by federally developed and administered tests whose results would be reported by federal officials. Although it is likely those tests would be administered annually, this need not be the case. Just as has been seen in NAEP assessment over the years, certain subjects could be assessed every few years rather than every year. If a series of more instructionally focused assessments were to be used in the "off years," then it may well be that periodic rather than every-year testing would be needed.

There is nothing in Control Option 5 to indicate that *all* curricular choices would be made at the federal level. Instead, federal accountability tests might focus only on certain pivotal subject areas such as language arts, mathematics, and science, and perhaps only at a few selected grade levels. (As noted above, were national tests to be installed, the magnitude of such a colossal assessment operation might well lead to the assessment of students' status at fewer

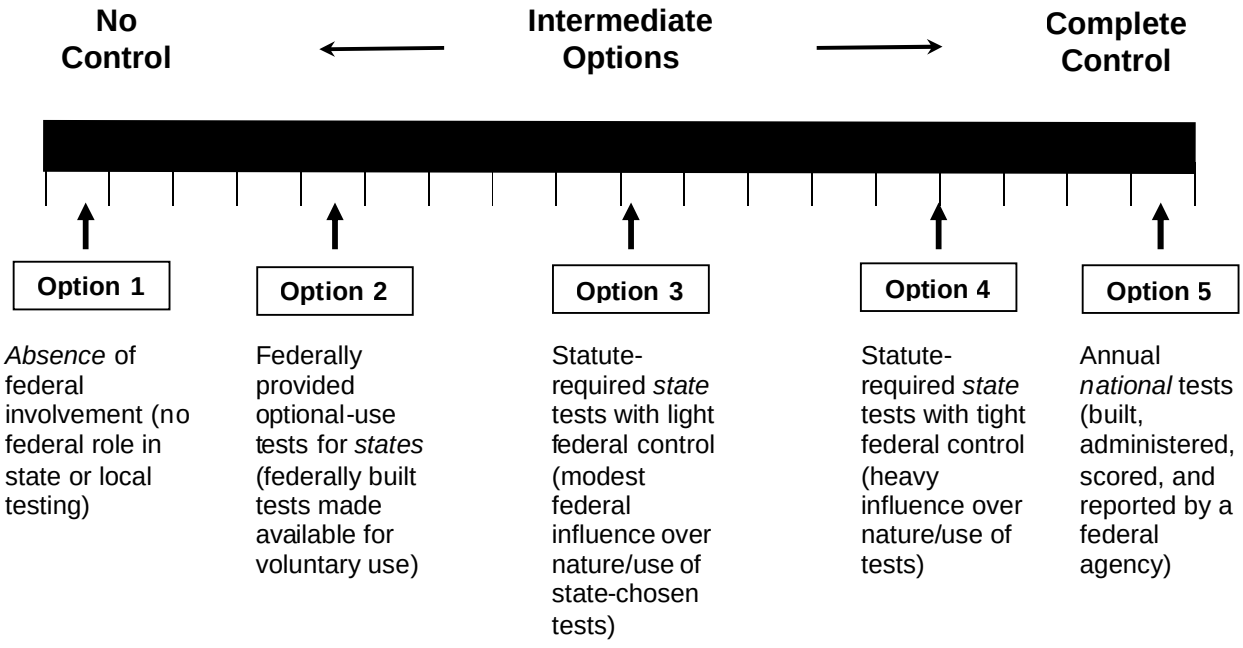
grade levels than are currently assessed because of NCLB, that is, in grades 3-8 and once in high school.) Thus, state and local education officials might still have responsibility for the promotion (and assessment) of curricular aims in certain subjects and at certain grade levels. In Control Option 5, however, the most pivotal accountability tests would be a federal responsibility.

Looking Back at the Control Options

In rethinking what might constitute an appropriate role for the federal government in relation to the kinds of accountability tests that are exerting an ever-increasing influence on what takes place in our schools, it seems apparent that the single most important factor to consider is *how much control over high-stakes tests the federal government should have.* There will surely be other issues to address as anyone tries to rethink what the best role is for the federal government when accountability tests are to be used. But the five control options portrayed graphically in **Figure 1** seem to represent the most viable alternatives facing those who are doing such rethinking.

When most of us are presented with a continuum of some sort from which we must make selections, our natural tendency is to avoid extremes and head, instead, toward an in-between stance. In the current situation, this might well be the best approach to take. But it might not be. Thus, *all* of the five alternatives represented in the control option range shown in Figure 1 should be given careful attention—including the two extreme control options, that is, the complete federal control of accountability tests or the complete absence of such control. Perhaps it may truly be time to jettison all federal control of state or local accountability tests.

Figure 1. A Range of Federal Control Options for Accountability Tests



Perhaps it may truly be time to abandon local or state control of accountability tests and to embrace, instead, a genuinely national approach to the evaluation of educational quality. Clearly, blends of one or more of these five control options can be used at different levels and for different subject areas or assessment functions. But a re*thinking* of a suitable stance for the federal government regarding our nation's accountability testing should always get under way by attempting to answer this overridingly important question: *How much control?*

THREE DESIGN DIMENSIONS FOR ACCOUNTABILITY TESTS

If the most important choice in rethinking a federal assessment role is the amount of control we should assign to federal authorities, then a close second-place choice is *what kinds of tests should be built?* Let's consider, then, three factors that should be addressed when designing any kind of accountability test. There are three principal missions an educational accountability test can be built to accomplish. These missions can be seen in educational tests irrespective of whether those tests are built by federal, state, or local authorities. But the greater the influence an educational accountability test has, the greater will be its impact on students. Thus, national accountability tests, or nationally influenced accountability tests, deserve even greater scrutiny with respect to these three missions than do state or local accountability tests.

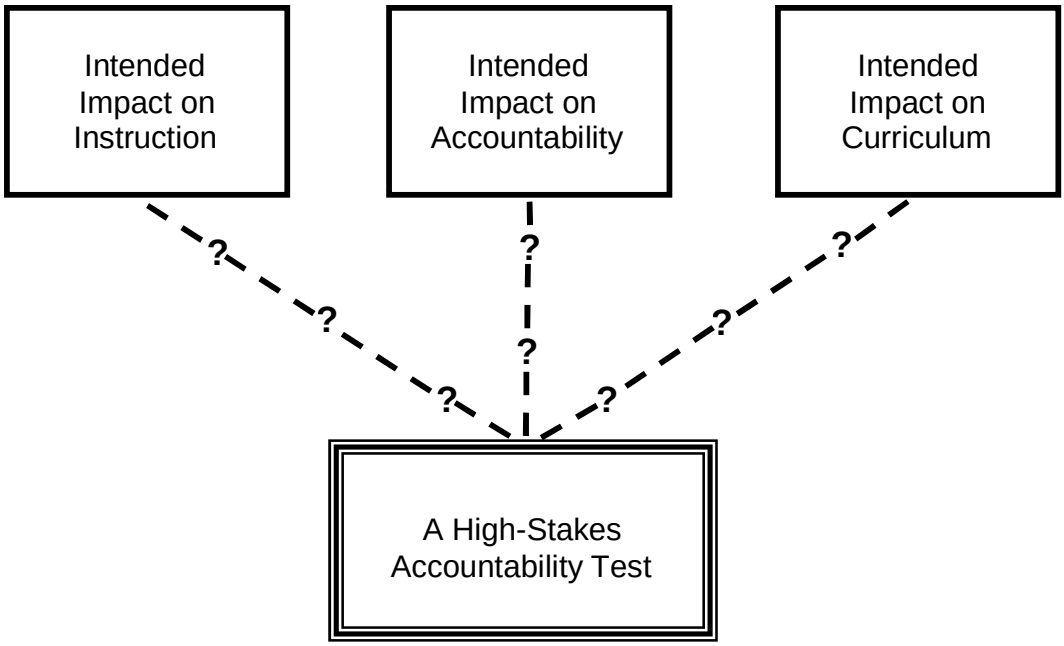
Accountability tests can be constructed in such a way so they are more likely to influence *instruction, accountability,* or *curriculum.* If the tests are built skillfully, odds are that they will accomplish their intended missions. Indeed, although difficult, it is even possible to craft an accountability test so that it simultaneously tackles two or three of these distinctive measurement missions.

Tests differ. That short sentence is more significant than it might appear to be. Sadly, many Americans wrongly believe that, "a test is a test is a test." Few assertions could be more flawed. Depending on how educational tests are built, those tests can fulfill markedly different functions. Going back as far as World War I, most U.S. educational tests—whether they were achievement tests or aptitude tests—have been constructed so they yielded *comparative* score interpretations for test-takers. It was Robert Glaser (1963) who first called our attention to the difference between traditional comparative testing, which he called *norm-referenced* testing (because students' scores were "referenced," that is, interpreted by being compared to a norm-group's scores) and *criterion-referenced* testing (wherein students' scores were interpretively "referenced" to a criterion behavior such as a clearly defined skill or body of knowledge). These two approaches to educational assessment, though they might seem somewhat similar to a novice, serve fundamentally different measurement functions. To accomplish those two functions successfully, one must build the two types of tests differently.

Thus, regardless of the degree of federal control of assessments that might be present in the future, a rethought federal role in educational assessment must pay heed to whether the tests being developed should be constructed so they are apt to have an impact on instruction, accountability, and/or curriculum. Let's look, then, at how a test should be designed if it is to maximize its impact on, in turn, instruction, accountability, or curriculum. The three design dimensions to be addressed are identified in **Figure 2**, where it can be seen, from the question marks, that there is uncertainty regarding the degree to which each of those design dimensions will play a prominent role in the construction of a high-stakes accountability test. There are certainly other factors to consider when setting out to build any important test, but for purposes

of rethinking a federal role in educational accountability assessments, the three dimensions to be treated in this analysis are, without question, the most significant.

Figure 2. Potential Design Dimensions for Accountability Tests



I intend to briefly consider each of the three design dimensions, first describing the nature of a particular design dimension and then indicating what test builders would need to do in order to successfully accomplish such a measurement mission. I apologize for addressing these key choice points with such brevity that my treatment of them may seem cavalier. I don't wish to suggest that this sort of test-development endeavor is easy or that it doesn't require the very best thinking from the very brightest test developers. Creating accountability tests that have a decent chance of accomplishing *any* significant measurement mission is both intellectually taxing and fraught with unforeseeable challenges.

Moreover, as you will see, all that a test developer can do is craft an assessment so that it is *intended* to have an impact on, say, a state's curriculum. Only when the tests are actually used can one know whether this intention has been realized. Yet, in recognition of the potent influence accountability tests have on what goes on in our schools, it would be senseless not to try to have those tests exert a positive influence on U.S. schooling. In turn, then, let's consider how one might create an accountability test so it has a beneficial impact on (1) instruction, (2) accountability, and (3) curriculum.

Intended Impact on Instruction

This initial design dimension refers to the likelihood that an accountability test will have a positive influence on the actual instruction received by students. Remember, most accountability tests will already have been fashioned with a particular measurement mission in mind. That is,

all accountability tests are intended to supply evidence—in the form of students' tests scores—allowing evaluative judgments to be made regarding the quality of instruction being offered in a state, district, or school. This evaluative function, of course, can be accomplished by a particular test with a greater or lesser degree of success.

However, for the sake of exposition, let's assume for the moment that our federal government suddenly became much more heavily involved in accountability testing, such as would be the case if certain of the previously described control options were to be implemented. Let us also assume that, either in the authorizing of federal legislation or later on, because of the preferences of some type of governing or advisory group, it was decided that in addition to its accountability function, a series of new federally developed or monitored assessments should be deliberately designed *to improve the quality of instruction* provided by U.S. teachers. Were this to be the case, then the architects of any such instructionally oriented accountability tests should make sure to adhere to the following test-construction suggestions. Space limitations for this paper preclude a detailed treatment of these recommended actions, but each of the following six suggestions could be addressed in greater detail if there were any interest in having an accountability test function as a catalyst for improved instruction.

The curriculum aims being assessed must be patently defensible. Because this kind of instructionally oriented accountability test is deliberately intended to alter instruction, and will often have a meaningful focus on students' mastering each of the curricular aims being assessed, it is extraordinarily important to structure the test around the measurement of educationally sound curricular aims. Thus, those curricular aims must be subjected to far more than the kind of routine, run-of-the-mill scrutiny accorded the curricular aims assessed by certain state-level accountability tests. Indeed, an atypically careful appraisal of the quality of any potentially assessable curricular aims is requisite so that instructionally oriented accountability tests truly measure students' status regarding commendable curricular aims.

Only a modest number of genuinely significant curricular aims should be assessed. Too many curricular targets turn out to be no targets at all. Presented with hundreds of curricular goals to achieve, even the most fervent of teachers often become overwhelmed by so many targets. For an accountability test to improve instructional quality, it is necessary for the test to present to teachers an intellectually manageable number of curricular targets, more like a dozen curricular aims rather than 50 or more such aims. Besides, an accountability test with an instructional-improvement mission must provide reasonably accurate indications of students' status regarding *each* assessed curricular aim so that teachers and students can take action to address any unachieved curricular aims.

Too many curricular aims to be assessed make it literally impossible to accurately determine students' mastery status for each curricular aim. (There is insufficient testing time to include enough items for each curricular aim.) Accordingly, a serious *prioritization* effort will be necessary so that the most significant curricular aims will definitely be measured by an instructionally oriented accountability test. The challenge for test developers, therefore, is to coalesce collections of lesser curricular aims under a few broader, but essentially homogeneous, curricular targets. An example of such coalescence can be seen when language arts specialists subsume a flock of composition subskills and several genres of writing under a student's "ability to write original compositions of various types."

If an instructionally oriented accountability test is well conceived, teachers should be able to promote students' mastery of what's measured more efficiently. Teachers will often find they

have instructional time left over to address other curricular aims that, although not measured by an accountability test, are still worth pursuing.
Although accountability tests can galvanize a teacher's resolve to teach well those things that are to be tested, when there are too many things being tested, then most teachers become overwhelmed by so many assessment targets. Teachers end up guessing, often incorrectly, about what is going to be assessed on an upcoming accountability test. After a few years of mistakenly failing to emphasize what's being tested, and lavishing instructional attention on what's not tested, many teachers simply abandon any effort to promote their students' mastery of the curricular aims supposedly being assessed. Such abandonment, of course, negates any accountability test's positive impact on instruction.

Clear, readily comprehensible descriptions of each assessed curricular aim should be made available. Teachers who know what they want their students to learn are obviously more likely to promote their students' achievement of those things than are teachers who have no idea about what it is they want their students to learn. In accord with this incontestable truth, one of the most potent payoffs of an accountability test that's built to improve instruction is it can markedly clarify a teacher's conception of what is to be taught. A major dividend of a well-designed accountability test, therefore, is that it can enhance a teacher's *clarity of intent,* that is, the accuracy with which the teacher comprehends what it is that the students are supposed to learn.

If accountability tests are accompanied by clear, easily understood descriptions of the skills or bodies of knowledge to be assessed, then teachers, students, and students' parents will have acquired better understandings about curricular intentions. The more clearly teachers understand where their students are heading, the better job those teachers can do in trying to help their students reach those curricular destinations. Students and their parents, as well as teachers, can profit from having access to a set of lucid descriptions regarding what's to be assessed by an accountability test. If teachers genuinely understand the skills and knowledge being assessed, then they can focus their instructional efforts not on getting students to correctly answer particular test items, but instead, on promoting students' mastery of the skills and knowledge those items assess. Students and parents can do the same.

Enough items should be used to ascertain each student's status regarding every assessed curricular aim. An instructionally oriented accountability test is intended to help students accomplish every curricular aim being measured. Not only should teachers know how all their students are doing with regard to each assessed curricular aim, but students and students' parents should also know. Teachers, if they discover that their students performed dismally on certain curricular aims, can modify how they tried (and failed) to teach those things. Students and their parents can also use such per-aim reports to let them know where it is that a student needs to do serious "catch-up" work.

The actual number of items per curricular aim will depend on the *grain-size* (that is, breadth) of the curricular aims being assessed. Broader curricular aims, crammed with content knowledge and collections of subskills, will obviously require more items per aim than will narrower curricular aims. The actual numbers of needed items can usually be estimated with reasonable accuracy by teachers who are conversant with the content involved and are familiar with children of the ages being assessed. What's being sought here is not a perfectly accurate, error-free fix on a student's per-aim mastery, but instead, the sort of reasonably accurate estimate of a student's mastery of each curricular aim that can help teachers, students, and parents take sensible, evidence-abetted actions aimed at improved learning.

A sufficient variety of items should be employed so teachers will direct their instruction toward students' generalizable mastery of skills and knowledge. We want our students to master skills and bodies of knowledge in a way that they can apply this mastery widely in a host of school and nonschool settings. What we do not want our students to do is master a skill *only as it happens to be measured* in a limited, circumscribed manner. Thus, to the extent possible, accountability tests intended to nurture improved instruction should contain not only a single method of tapping students' mastery of curricular aims, but instead should incorporate a variety of measurement tactics.

Clearly, cost constraints are likely to prevent us from employing as many constructed-response items (for instance, short-answer and essay items) as we might like, but if there can be at least a small number of such items on a test, then teachers will be inclined to make sure their students have mastered a skill so those students can apply their skill mastery in a more generalizable manner. And even with the use of selected-response items such as multiple-choice or binary-choice items, variety in test items should be sought rather than the use of a single, eminently predictable way of measuring a student's mastery of a curricular aim. We want widely applicable mastery for our students, not narrow, test-constrained forms of mastery.

As many instructionally sensitive items as possible should be incorporated in tests. If the measurement mission of an instructionally oriented accountability test is to improve instruction, it is apparent that such a test should be able to distinguish between effectively taught and ineffectively taught students. Recent appraisals of the items used on certain accountability tests suggest that many of those items may be unable to differentiate between well-taught and poorly-taught students (Popham, 2008a). Yet, if teachers are going to be stimulated by an accountability test to do a better instructional job with their students, then any accountability test being used to evaluate teachers' success must be able to tell which instruction has been successful and which instruction hasn't. Recent attention to this issue suggests that it may be possible to employ both judgmental and empirical procedures to the determination of whether particular items on a high-stakes test are instructionally sensitive (Popham and Berliner, 2008).

If teachers' improved instruction fails to be transformed into improved test scores for their students, how long will those teachers rely on the results of such an accountability test to spur their instructional improvement efforts? Clearly, they'll not be using for long any accountability test that's instructionally insensitive. Although our technical base for gauging the instructional sensitivity of an accountability test is, at the moment, fairly primitive, reasonable attention should be given to ensuring that a substantial number of the items on an instructionally oriented accountability test will be likely to distinguish between high-quality and low-quality instruction.

Summing up, then, I have identified six suggested actions that should be considered by those who are attempting to create an accountability test purporting to have a positive impact on instruction. Whether such instructionally oriented accountability tests are employed as a component of one or more of the several control options previously described, these six suggestions still pertain. I cannot claim that there is a minimum number of the six suggestions that must be followed if an accountability test is going to truly have a positive influence on instruction, but I do contend that if all or most of the six suggestions associated with this particular design dimension are disregarded, the resulting likelihood of an accountability test's improving instruction will be negligible or nonexistent.

Intended Impact on Accountability

Generally speaking, most of the state and district tests currently used in the nation for school evaluation have been constructed with accountability exclusively in mind. This second of our three design dimensions deals with the likelihood that students' performances on an accountability test—whether a national, state, or local one—will play a prominent role in the appraisal of educators' success. Accordingly, if this particular design dimension is chosen to guide the construction of an accountability test, as was true with the previously described design dimension, there are several experience-based suggestions to be proffered that, if followed, are likely to maximize the contribution of such tests to the decisions that must be made as part of an accountability program. Four suggestions are presented below.

Test-score evidence should be provided as a decision-relevant unit of evaluative analysis. Although we currently find few accountability tests deviating from this initial suggestion, it is worth a reminder that decision-makers at some identifiable action-taking level—such as at a school, a school district, or the schools of an entire state—will need to rely on the results of accountability tests. Thus, if a federal statute is aimed at holding *states* accountable for their state-level implementations of federally supported programs, then it is obvious that state-specific test results must be provided. Similarly, if states intend to hold districts or schools accountable by evaluating district or school test-based success, then test-score evidence must be generated so that test-based reports and, thereafter, evaluation-dependent decisions, can be made at both the district level and the school level.

Results on accountability tests should be readily understandable to all relevant audiences. There was a time, eons ago, when students' test scores were reported to students and their parents almost uniformly as "number correct" or, more commonly, as "percent correct." People understood what "number correct" meant. People understood what "percent correct" meant. Now we increasingly find that students' results on accountability tests are reported in the form of abstruse scale-scores derived from the unfathomable machinations of computers using programs apparently designed to preclude comprehension by humans. If it becomes genuinely *impossible* to make students' performances on an accountability tests comprehensible to those who must rely on this evidence, then perhaps other tests should be used. In some instances, of course, too-simple reporting mechanisms (such as "percent correct") might misrepresent students' test results. But if a sophisticated item-response theory approach, along the lines of an exotic three-parameter analysis, does a remarkably accurate job of representing a set of test performances, yet that analytic approach can't be understood by the very people who must use the test's results, then what good are those results?

This suggestion, namely, that the reported results of accountability tests must be understandable to concerned constituencies, does not preclude the use of sophisticated statistical analyses for other necessary psychometric functions, such as the year-to-year equating of different tests. Nonetheless, *after* such statistical machinations have been accomplished, the results of accountability tests must be reported in ways that can be readily comprehended.

Results of all tests should be regarded as patently credible, that is, as trustworthy indicators of educators' instructional successes. In recent years we have seen frequent newspaper "exposés" in which performances on state accountability tests are contrasted with those of NAEP, only to have reporters indicate, with apparent incredulity, that far fewer students are deemed proficient on NAEP (a national test) than on a state's NCLB tests. When this happens, of course, it is widely perceived that the state's tests are "soft," "self-serving," or

"undependable." Well, test results seen to be soft, self-serving, or undependable rarely possess sufficient leverage to exert a positive influence on the decisions made regarding our schools.

In the contrasts between NAEP and state results, there are many factors contributing to the disparate numbers of students classified as "basic," "proficient," and "advanced." Among these "state versus NAEP" factors are the cut scores that have been set to distinguish among students in different categories, the difficulty of the tests themselves, and the particular skills and knowledge being measured by each test. In some instances, to be candid, there is some reason to believe that state officials may have established indefensibly low cut scores for students' performance categories simply to avoid the perception that their state's schools were unsuccessful. But in many settings, there are other considerations, most of them legitimate, that could have led to NAEP-versus-state inconsistencies.

What must be recognized is that if any accountability test, whether it be a local, state, or national one, is seen by the citizenry to be lacking credibility, then its results are essentially useless. Either a serious educative campaign should be undertaken to help people see why a test is credible or, if this is impossible, the test should be replaced with a more credible one. Given the cut scores adopted for NAEP categories, it may be necessary to argue that NAEP's "basic" level of performance is the one that should be contrasted with most states' "proficient" level performances on their own state accountability tests. But if such an argument is undertaken, it should be accompanied by ample support.

Any test must be able to distinguish between effectively taught and ineffectively taught students. The fundamental premise of most accountability programs is that students' test results will reveal which educators have been doing a good job (so, if possible, those educators can be rewarded) and which educators have not been doing a good job (so actions can be taken to improve what those educators have been up to). What this premise requires, in order to be satisfied, is the use of tests that can accurately distinguish between successfully taught and unsuccessfully taught students.

As noted earlier, we are just beginning to develop technical procedures for scrutinizing the instructional sensitivity of items on accountability tests. However, drawing on whatever analytic procedures seem currently defensible, it is imperative that accountability tests be able to help decision-makers distinguish between successfully and unsuccessfully taught students. If accountability tests fail to do this, then these tests are fundamentally unfair to the educators whose instructional prowess is being appraised on the basis of their students' test scores.

In review then, we have considered four suggestions that should be attended to if one is relying on this second design-dimension associated with the development of an accountability test, that is, to build accountability tests that can be used to accurately appraise educators' success. As pointed out, we have more experience in using accountability tests for this purpose because, frankly, this has been the dominant use of such tests during the last few decades. The following design dimension regarding curriculum, in contrast, has been much less studied. We turn, then, to the final of our three design dimensions, an accountability test's intended impact on the curriculum.

Intended Impact on Curriculum

Because what's assessed by accountability tests can have such a profound impact on what gets taught, there is also another potential role for accountability assessments, namely, to alter the actual curricular aims being pursued by educators. To illustrate:If a state's annual accountability

tests were modified so that they began assessing students' mastery of "decision-making skills," we can be certain that in many classrooms we would soon find substantial attention being given to students' acquisition of such decision-making skills. Because of the numerous social, cultural, and technological changes in today's world, it has been argued with increasing vigor that many of yesteryear's curricular aims should be replaced with what are often described as "21st century skills." (Regan, 2008) If someone were to support such a curricular goal-switching strategy, one effective way of doing so would obviously be to require the assessment of 21st century skills by the nation's accountability tests.

This final design dimension dealing with an accountability test's impact on curriculum seems, in a sense, to be somewhat out of sequence. That is, we are usually told that an accountability test should be constructed so it will assess students' attainment of whatever curricular aims have been chosen for the schools. However, even though focusing on an assessment prior to a set of curricular aims may border on what seems to be another variation of the classic chicken-or-egg conundrum, there is little doubt that whatever curricular aims federal authorities urge be included on our accountability assessments will be pursued in our schools. If an accountability test is going to be created in an attempt to foster instructional attention to extant or, especially, new curricular aims, the following four suggestions should be considered.

Any innovative curricular aims to be assessed should command the support of substantial numbers of stakeholders. Time confers legitimacy, or so it sometime seems. The longer something has been in place, the more appropriate that thing is thought to be. Such is surely the case with the curricular aims for our schools. American educators have been pursuing curricular aims for years that, in many instances, seem to have been in place almost since our nation's birth. Thus, any movement toward the infusion of innovative curricular aims is certain to encounter at least some resistance from those for whom "time-tested" is regarded somehow as "time-proven."

This predictable resistance to new curricular aims suggests that any innovative curricular aim slated to be assessed by an accountability test must command the approval of a substantial number of supporters. The assessment of new curricular aims dare not be seen as a bizarre idea applauded by only a few proponents. If this is the case, it is unlikely there will be sufficient support for such assessment, so students' mastery of the new curricular aims will ever be measured. The installation of atypical curricular aims on an accountability assessment, however, need not be *universally* endorsed. Nor need there be a *dominant consensus* in favor or the newly assessed curricular targets. But there should at least be a substantial number of advocates of these new curricular aims—advocates who can spell out with cogency the reasons these new aims ought to be assessed.

Innovative curricular aims to be assessed should be few in number. Historically, when educators set out to identify the curricular targets at which a school's energy should be directed, they end up identifying too many such targets. As a consequence, teachers are often presented with a plethora of curricular aims and, therefore, are unable to provide truly in-depth instructional treatment of many, if not all, of those aims. Precisely the same problem will arise if the proponents of curricular change attempt to assess students' mastery of too many innovative curricular targets. If attention to this design dimension is truly intended to infuse new curricular foci into our schools, then having too many innovative curricular targets is certain to deflect instructional treatment from those targets.

This is no time for those who want the schools to pursue new curricular aims to assemble a "wish list" of *all* the new skills and knowledge tomorrow's adults should master. It makes more

sense to infuse gradually a set of, say, three or four powerful new curricular aims and see those aims successfully pursued than it does to seek students' achievement of 20 or 30 innovative aims and see this excessive array of curricular aims overwhelm teachers. Because this is another clear instance in which less turns out to be much more than more, care should be taken to build accountability tests measuring the most significant of a thoughtfully *prioritized* set of innovative curricular aims. Care must be taken, of course, to avoid adoption of curricular aims embodying grain sizes so large that what is being sought of students is masked by the excessive generality of a curricular aim.

Practical, cost-effective ways to assess students' mastery of any new curricular aims should be at hand. Because the crux of this design dimension rests on our ability to create accountability tests measuring innovative curricular aims, it will be immediately necessary to lay out the assessment tactics (hopefully identifying more than one such assessment tactic) to measure students' accomplishment of each new curricular aim to be assessed. Ideally, as a vehicle for clarifying the meaning of any innovative curricular aim to be measured on an accountability test, several of these assessment tactics should be described along with illustrative sample items showing how students' attainment of a curricular aim is likely to be measured.

Ancillary support materials should accompany an accountability test assessing atypical curricular aims. One thrust of an accountability test intended to measure innovative curricular aims should be to have teachers focus some of their instructional attention on the new curricular targets and thus get students to master whatever has been set forth in the new curricular aims. But most of these innovative curricular aims are apt to be precisely what they say they are, namely, *innovative.* If curricular aims are genuinely new, then it is likely teachers will need some support to be successful in promoting their students' accomplishment of these unfamiliar curricular aims—and *time* to become adept at getting students to achieve those aims.

And this is why, in order for this third design dimension to become effective, total reliance cannot be placed on the test itself. The test in isolation, especially because atypical curricular targets are being assessed, is unlikely to accomplish serious shifts in classroom practices. Accordingly, accompanying the test should be descriptions of alternative instructional procedures (surely more than one) that teachers might adopt if they wish their students to master the new to-be-assessed curricular aims. Guidelines could be provided regarding instructional ploys to be considered as well as the kinds of en route tests that might be employed when a classroom formative assessment process is being used to promote students' mastery of the new curricular aims. Indeed, suggestions could also be proffered regarding the kinds of "building block" subskills and enabling knowledge to be achieved by students as they move toward mastery of an innovative curricular aim. In general, the creators of these kinds of accountability assessments must try to make it easier for teachers to successfully promote students' mastery of any unfamiliar curricular aims that will be measured by a reformulated accountability test.

There's another issue that needs to be briefly addressed whenever the potential curricular impact of an accountability test is to be discussed, and the issue is *curricular reductionism.* Many educators fear that if so much instructional attention is given by teachers to promoting students' mastery of what is to be measured by accountability tests, then other important subjects and curricular aims will be left in the dust. Ideally, if accountability tests were well conceived, it would be possible to promote students' mastery of what's measured with sufficient *efficiency* that there would be instructional time left over to pursue students' achievement of

other important curricular aims, even though such aims were not assessed by a state's accountability tests.

Realistically, however, to minimize the likelihood of curricular reductionism, it may be necessary to devise accountability tests so that such subjects as social studies, health education, or visual arts are assessed on a sampling basis, that is, whereby only the students in certain districts or certain schools are asked to complete separate, unpredictable segments of an accountability test measuring mastery of skills and content areas the state's education authorities do not wish to see seriously de-emphasized by the state's teachers. Because these curricular aims are *eligible* to be tested, even though most students will, in fact, not be assessed with such items in a given school year, teachers will be loath to completely abandon instructional treatment of the content not included in the mainline accountability tests.

As an incentive to get educators to attend to a full range of curricular aims, public reports could be issued regarding the performance of those schools and districts chosen to take part in such sample-based segments of a state's accountability assessments. If this approach proves insufficiently potent in combating curricular reductionism, then a state might promulgate directives requiring that minimum proportions of instructional time must be devoted to content areas not covered by accountability tests. The state education agency then would surely need to employ some sort of audit mechanism, such as an inspectorate system whereby occasional, unannounced classroom visitors made sure that sufficient segments of instructional time were being devoted to curricular aims other than those measured by the state's accountability tests.

In review, then, if this third design dimension were adopted for use in the creation of an accountability test, care would have to be taken so that the tests themselves, as well as any associated materials dispensed along with them, would have a viable chance of making a difference in pursuit of the curricular aims being assessed.

As suggested earlier, these three design dimensions might be employed in concert or separately. Clearly, if an accountability test sets out to simultaneously bring about improvements in instruction, accountability, and curriculum, the test-makers' tasks are going to be meaningfully more challenging than if only one measurement mission at a time were being sought. Nonetheless, given sufficient thought and planning, it would definitely be possible to build accountability tests incorporating all three of the design dimensions.

LOOKING BACK

Rethinking how any federal government ought to take part, if at all, in the educational assessment of its nation's students is a consummately complicated undertaking. But if we are serious about doing so, then we dare not become so enmeshed in the complexities of this endeavor that we end up failing to arrive at any sensible strategy about how to proceed. *Simplicity* is a good thing. It can describe a sometimes complex phenomenon in a way that is easy to understand. *Simplism*, on the other hand, represents excessive simplification. In any rethink of how the federal government should interact with the nation's most significant educational tests, we need to aim for the simple and eschew the simplistic.

In looking back at the issues associated with how we rethink a federal role in educational assessment, I have argued that the two most important questions to confront are (1) *How much control should we have the federal government exercise over the nation's accountability tests,* and (2) *What should be the measurement mission(s) of those tests?* If these are, in fact, the two

most important questions, then an appropriate framework for considering federal involvement in educational testing should rest on those two fundamental axes, namely, *degree of federal control over tests* and *the intended measurement mission(s) of those tests.* If we confront those two issues first, we will be tackling our task simply—by focusing on what's truly most significant. This might seem simplistic to some, but if we don't commence our deliberations by answering these two overridingly important questions, we're unlikely to make much progress at all.

Within the context of such a two-factor framework, we will still be obliged to deal with a series of related and remarkably thorny issues—for example, how to establish acceptable cut scores for whatever categories we finally choose to represent levels of student achievement. Then there is the choice between the kind of status-focused evaluative model currently imbedded in NCLB and some approach more attentive to students' improved achievements. Many such problems must be solved. The battleground of U.S. educational accountability is littered with more than a few abandoned ideas thought worthy by their originators. There is clearly more to educational accountability programs than accountability tests. Fine accountability tests do not necessarily lead to fine accountability programs. Yet, because it appears that for the foreseeable future we will be relying on *test-based* educational accountability, the weaker our accountability tests, the more woeful will be our accountability systems.

Let's look back, then, on what has been going on for about 50 years with regard to federal laws and the two major issues treated in this paper, that is, degree of federal control and the nature of the measurement mission(s) to be pursued by educational accountability tests.

Regarding level of federal influence, we have seen—chiefly through the various incarnations of ESEA—an ever-expanding federal footprint on educational accountability tests. Not only have the stakes been bumped higher, but the number of assessed students has increased and the constraints on the nature of the tests have been intensified. Whether this is an irreversible trend remains to be seen. In the early days of ESEA, there were surely provisions in federal law regarding educational accountability tests, but the impact of those provisions on U.S. accountability tests was, at best, modest. Now, with the most recent version of ESEA, that is, NCLB, we find a remarkable tightening of federal impact on the nation's educational accountability tests and, as a consequence, what goes on in the nation's schools.

Turning to the second major choice point for educational accountability tests —what we want them to accomplish—for more than four decades we have seen a fairly clear emphasis in all federally influenced accountability tests. The *only* measurement mission of these tests has been that of accountability. The tests have not been designed to improve instruction or to install innovative curricular aims. Note I am not arguing that *laws,* such as NCLB and IDEA, have failed to have an impact on instruction or curriculum. Clearly, they have. For instance, NCLB's obligatory disaggregation of student subgroups has brought an unprecedented level of instructional attention to those student subgroups. What I am suggesting, rather, is that the tests *per se* were not intended to have an instructional or a curricular impact. But, of course, if the nation's accountability tests are deliberately conceptualized to have either an instructional impact or a curricular impact on schooling, then they will have a markedly better chance of doing so. To make sure that such measurement missions are achieved, of course, ongoing evaluative attention to the real-world impact of those accountability tests will be needed.

It may have been noted that throughout the foregoing analysis I have not taken a position of advocacy regarding either of these two key questions, that is, an appropriate level of federal control or the right measurement mission(s) for important accountability tests. But if there are to be any sort of serious deliberations regarding how our federal government should tackle

educational assessments, I am convinced that we must *first* dig hard into the virtues of each of the various control options and, *immediately thereafter*, into the merits of the three design dimensions. Once we have adroitly wrestled these two problems to the mat, our remaining tasks might, comparatively, turn out to be a piece of cake—chocolate, of course.

AN ADVOCACY POSTSCRIPT

In this paper I tried to analyze as evenhandedly as I could the two critical issues of a best level of federal control for key educational tests and the measurement missions such tests should attempt to fulfill. In this postscript, I offer my own preferences for what I think would be (1) an appropriate level of control for federally influenced assessments and (2) suitable measurement missions for such assessments. It is tempting to undergird my position with a flock of forensic support, but postscripts are supposed to be brief, so I shall be.

With respect to the five control options, the one I think is most defensible is Control Option 4, that is, statute-required state tests with tight federal control. Yes, this option essentially mirrors the federal control we currently see in NCLB, and NCLB has surely been on the receiving end of sometimes scathing criticism. Yet, I regard Control Option 5, total federal control, as representing an approach unlikely to garner sufficient approval from those who must approve it if educational assessment is ever going to improve U.S. students' learning. And Control Option 3, where there is only light federal control, has already been shown under IASA to be ineffectual. I realize that ESEA is slated for reauthorization relatively soon, and I am convinced that if two serious shortcomings in NCLB could be rectified, then an improved version of this law could be educationally beneficial.

First, and most important, the targets for improved student achievement must be made realistic, not patently unattainable. The unrealistic target of having 100 percent of students reach proficient or better achievement levels by 2014 is causing NCLB to impale itself on its own unrealistic sword. To accomplish this kind of change in expectations without the political embarrassment of "softened standards" that might accompany it, I believe we must switch to a new way of reporting students' achievement levels, for instance, by using a set of labels built around students' being able to perform "at or above grade level." Parents and others know, at least generally, what "grade level" represents. And if we set those grade level designations in a defensible manner (attending not only to what our students currently can do, but also to what they *should* be able to do), then we could employ this new way of categorizing how our students are performing. Moreover, at the moment we shift the labels being used to report students' status, we could simultaneously establish a more realistic aspiration regarding how many of our students will be able to reach or exceed "grade level" at a given point in time. There is nothing inherently absurd in wanting less than 100 percent of our students to achieve or exceed grade-level performance. In England, for example, a governmentally established target, set some years ago, of having 85 percent of students achieve a desired level of performance has yet to be achieved. Great educational harm has been caused by NCLB's incorporation of an infeasible achievement target.

A second shortcoming of NCLB that would need to be remedied if a sensible implementation of Control Option 4 were to be installed deals with the tests that states are currently employing to evaluate their states' educators. In the early days of NCLB's implementation, essentially no meaningful guidance was given by federal officials to state assessment personnel regarding appropriate NCLB accountability tests, and as a result, almost all states plowed ahead with reincarnations of the assessments they had been previously using to

satisfy IASA. Those tests were, for the most part, altogether inconsistent with a federal law whose accountability cornerstone required educators to be evaluated on the basis of their students' improved scores on NCLB-approved accountability tests. When the U.S. Department of Education finally got around to providing assessment guidance to states, most states were locked into the use of tests that were, in my view, fundamentally incapable of detecting improved instruction even if such improved instruction had been present. Beyond that, the belated federal guidance proffered to states, especially via the peer review process, embodied the kind of traditional psychometric thinking that, in meaningful ways, is antithetical to an assessment-based accountability law predicated on using assessments not only to evaluate schools on the basis of instructional improvement, but also to help stimulate such improvements.

To help states create the kinds of tests called for in a system where tight federal control is operative, a substantial professional development and technical assistance effort at the federal level would be necessary so states would receive ample support in adopting new assessment devices that were compatible with a more instructionally oriented version of ESEA. In that regard, one possibility to consider would definitely be the federal provision of optional assessment instruments as suggested in Control Option 2. Helping state assessment personnel come up with more suitable assessment approaches can be promoted federally via professional development, supplying technical assistance, and providing exemplars of first-rate assessments. Such support would surely be requisite if Control Option 4 is to succeed.

Accordingly, with more realistic achievement targets and more appropriate assessments to use when measuring students' progress toward those targets, a successor to NCLB based on Control Option 4 could definitely promote the improvements in America's educational system that we all desire.

I turn now to the second of the paper's two issues, that is, what kinds of measurement missions should be undertaken by the federal government, and I am assuming that some variant of Control Option 4 would be in place. It will be recalled that, for the three design dimensions considered in the paper, significant tests could be built so they were intended to have an impact on instruction, accountability, and/or curriculum. Although heretofore we have seen federally influenced tests created chiefly to play a role in accountability programs, that is, to evaluate the efforts of educators as they operate schools and school districts, I would argue that *all three measurement missions* should be accomplished by the kinds of state accountability tests foreseen in Control Option 4. Yes, I would want these federally influenced tests, perhaps influenced by a system that procedurally was not dramatically different from the current peer-review process, to deliberately set out to have statewide assessments (1) serve as catalysts for improved instruction, (2) supply the evidence needed to make defensible evaluative decisions regarding the instructional effectiveness of schools/districts, and (3) promote adoption of the most appropriate curricular aims for today's and tomorrow's demands.

Can this sort of tripartite measurement mission succeed? Based on my personal experience a decade ago as a developer of high-stakes achievement tests, I am certain that this measurement triple-play can be accomplished. It will not be easy, but it is definitely doable.

Summing up this postscript, I advocate the adoption of a system in which a federal law calls for the provision of significant educational assessments at the state level, but those assessments should then be heavily influenced by federal officials via what should be a

transparent, independently monitored guidance system. With respect to the design dimensions to be incorporated in those state tests, the federal government should push state tests to be constructed to they have a positive impact on instruction, accountability, and curriculum.

Could such a system function in a way that would eliminate today's antipathy toward NCLB and, at the same time, markedly benefit the children in our schools? I believe it would.

REFERENCES

Anderson, L. (April 4, 2005). The No Child Left Behind Act and the legacy of federal aid to education. *Education Policy Analysis Archives, 13*(24). Retrieved July 27, 2008 from http://epaa.asu.edu/epaa/v13.n4/

Anderson, L. (2007). *Congress and the classroom: From the cold war to No Child Left Behind.* University Park, PA: Pennsylvania State University Press.

Borman, G. D. (2000). Title I: The evolving research base. *Journal of Education for Students Placed At Risk, 5*(1 & 2): 27-45.

Borman, G. D., & J. V. D'Agostino (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis, 18*(4): 309-326.

Borman, G. D., & J. V. D'Agostino (2001). Title I and student achievement: A quantitative synthesis. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.). *Title I: Compensatory education at the crossroads* (pp. 25-58). Mahwah, NJ: Lawrence Erlbaum Associates.

Bowers, J. J. (1991). Evaluating testing programs at the state and local levels. *Theory into Practice, 30*(1), 52-60.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (1999). *NAEP 1999: Trends in academic performance: Three decades of student performance*. (NCES Rpt. No. 2000-469). Retrieved September 8, 2008 from http://nces.ed.gov

Center on Education Policy. (2008). *Has student achievement increased since 2002?* Washington, DC: Author.

The Commission on No Child Left Behind. (2007). *Beyond NCLB: Fulfilling the promise to our nation.* Retrieved August 10, 2008 from http://www.aspeninstitute.org

DeBray, E. H. (2005). Chapter 2: Partisanship and ideology in the ESEA reauthorization in the 106th and 107th Congresses: Foundations for the new political landscape of federal education policy. *Review of Research in Education, 29*, 29–50. Retrieved July 27, 2008 from http://rre.aera.net

DeStefano, L. (1992). Evaluating effectiveness: A comparison of federal expectations and local capabilities for evaluation among federally funded model demonstration programs. *Educational Evaluation and Policy Analysis*, *14*(2), 157-169.

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and U.S. Department of Education reviews under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State School Officers. Retrieved August 23, 2008 from http://www.ccsso.org/publications/details.cfm?PublicationID=215

Fast, E., & Erpenbach, W. J. (2004). *Revisiting statewide educational accountability under NCLB: A summary of state requests in 2003-04 for amendments to state accountability plans.* Washington, DC: Council of Chief State School Officers. Retrieved August 23, 2008 from http://www.ccsso.org/publications/details.cfm?PublicationID=215

Fuller, B., Wright, J., Gesicki, K., & Kang E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher*, *36*(5), 268-278.

Gaddy, B., McNulty, B., & Waters, T. (2002). The reauthorization of the Individual with Disabilities Education Act: Moving toward a more unified system. [policy brief]. Aurora, CO: Mid-continent Research for Education and Learning.

Gardner, Walt (2008). The case for national standards and testing. *Education Week,* a Commentary essay scheduled for inclusion in the September 17, 2008 issue.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes—some questions. *American Psychologist,* 18: 519-21.

Hager, L. D. & Slocum, T. A. (2002). Alternate assessment: No Child Left Behind during statewide testing. *Rural Special Education Quarterly, 24*(1*),* 54-59.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2008). *Standards-Based Reform in the United States: History, Research, and Future Directions,* Washington, DC: Center on Education Policy.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-327.

Herman, J. L. (2007). *Accountability and assessment: Is public interest in K-12 education being served?* (CRESST Tech. Rep. No.728). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hombo, C. M. (2003). NAEP and No Child Left Behind: Technical challenges and practical solutions. *Theory into Practice*, *42*(1), 59-65.

Itkonen, T. (2007). PL 94-142: Policy, evolution, and landscape shift. *Issues in Teacher Education*, *16*(2), 7-17.

Keogh, B. K. (2007). Celebrating PL 94-142: The Education of All Handicapped Children Act of 1975. *Issues in Teacher Education*, *16*(2), 65-69.

Jennings, J. F. (2001). Title I: Its legislative history and its promise. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.). *Title I: Compensatory education at the crossroads* (pp. 1-24). Mahwah, NJ: Lawrence Erlbaum Associates.

Lehr, C., & Thurlow, M. (2003). *Putting it all together: Including students with disabilities in assessment and accountability systems* (Policy Directions No. 16). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved September 4, 2008 from http://education.umn.edu/NCEO/OnlinePubs/Policy16.htm

McDonnell, L. M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education*, 80(2), 19-38.

McLaughlin, M. J., & Nagle, K. M. (2004). Leaving NO child behind: Accountability reform and students with disabilities. In S. Mathison and E. W. Ross (Eds.), *The nature and limits of standards based reform and assessment*. NY: Praeger.

Merkel-Keller, C. (April, 1986). The evolution of evaluation: Title I to Chapter I. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mills, J. I. (2008). A legislative overview of No Child Left Behind. *New Directions for Evaluation, 2008*(117), 9-20.

National Center for Educational Statistics, (2006). *The state assessment: About state NAEP.* Retrieved August 10, 2008 from http://nces.ed.gov/nationsreportcard/about/state.asp

Popham, W. J. (2008a). *An open letter to state assessment directors* (regarding the instructional sensitivity of accountability tests). A May 1, 2008 e-mail transmission to all U. S. state assessment directors.

Popham, W. J. (2008b). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J., and D. Berliner (March 24-28, 2008). Empirically determining the instructional sensitivity of an accountability test. Paper presented at the annual meeting of the American Educational Research Association, New York.

Redfield, D., & Sheinker, J. (2004). *Framework for transitioning from IASA to NCLB.* Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers. Retrieved August 23, 2008 from http://www.ccsso.org/publications/details.cfm?PublicationID=215

Regan, Bob (July/August, 2008). Why we need to teach 21st century skills—and how to do it. *MultiMedia & Internet @ Schools Magazine*.

Riddle, W. C. (1989). *Education for disadvantaged children: Major themes in the 1988 reauthorization of Chapter 1. (*89–7EPW). Washington, DC: Congressional Research Service.

Riddle, W. C. (1998). *National assessment of educational progress: Background and reauthorization and issues (*98–348 EPW). Washington, DC: Congressional Research Service.

Riddle, W. C. (2001). *Assessment requirements under ESEA Title I: Implementation status and issues* (RL30742). Washington DC: Congressional Research Service.

Riddle, W. C. (2008). *Educational testing: Implementation of ESEA Title I-A requirements under the "No Child Left Behind Act"* (RL31407). Washington, DC: Congressional Research Service.

Rutherford, W. L., & Hoffman, J. V. (1981). Toward implementation of the ESEA Title I evaluation and reporting system: A concerns analysis. *Educational Evaluation and Policy Analysis*, *3*(4), 17-23.

Schlechty, Phillip (2008). No community left behind. *Phi Delta Kappan, 89*(8), 552-559.

Shaul, M. S., & Ganson, H. C. (2005). The No Child Left Behind Act of 2001: The federal government's role in strengthening accountability for student performance. *Review of Research in Education, 29*, 151-165. Retrieved July 27, 2008 from http://rre.aera.net

Spring, J. (2008). *The American school: From the Puritans to No Child Left Behind.* New York: McGraw Hill.

Stake, R. E. (2007). NAEP, Report cards and education: A review essay. *Education Review, 10*(1), Retrieved from http://edrev.asu.edu/essays/v10n1index.html

Thomas, J. Y., & Brady, K. P. (2005). The Elementary and Secondary Education Act at 40: Equity, accountability, and the evolving federal role in public education. [Electronic version]. *Review of Research in Education*, *29*, 51-68.

Thurlow, M. L., Lazarus, S. S. Thompson, S. J., & Morse, A. B. (2005). State policies on assessment participation and accommodations for students with disabilities. *The Journal of Special Education, 38*(4), 232-240.

U.S. Department of Education Office of Special Education Programs. (2007). *IDEA regulations. Alignment with the No Child Left Behind (NCLB) Act.* Retrieved August 23, 2008 from http://idea.ed.gov/explore/view/p/%2Croot%2Cdynamic%2CTopicalBrief%2C3%2C

Zettel, J., & Ballard, J. (August, 1979). The Education for All Handicapped Children Act of 1975 PL 94-142: Its history, origins, and concepts. *Journal of Education*, *161*(3), 5-22.