Beyond Standardization in School Accountability

Mindy L. Kornhaber

Pennsylvania State University
October 2006

Prepared for the Measurement and Accountability Roundtable
Washington, D.C. November 16-17, 2006

DRAFT: November 7, 2006

Any system of educational accountability in the United States should be guided by two essential, and equally important, aims. First, it should be informative: It should allow the public to know the status and progress of students' knowledge, skills, and understanding within and across schools. Second, the system should be cognitively constructive: It should advance all students' learning and enable educators to improve instruction (see Fredericksen & Collins, 1989).

Neither of these aims is being met under the No Child Left Behind Act, as Part I of this paper explains. Part II uses the groundwork of informative and cognitively constructive aims to eliminate ambiguities in the meaning of multiple measures. Such ambiguities have impeded substantive uses of multiple indicators. Part III describes the components of a system of multiple measures. This system would not be simple or easy to build, but it stands a far greater chance of producing genuine benefits across a wide range of students than a system based on a single high-stakes test. Part IV provides some policy recommendations that may be useful in revising NCLB, so that it can begin to address informative and cognitively constructive aims.

Part I: Failure to Meet Informative and Constructive Aims Under NCLB

A. Deficiencies in Meeting Informative Aims:

Our ability to obtain reasonable information from standardized testing rests on several assumptions. One is that the tests have been administered under the same or similar conditions. Given standardized test materials and procedures, it becomes possible to say that test scores shed light on the tested topic rather than matters extraneous to it. Of course, standardization is never perfect. The question is, how far from that ideal is it possible to roam and still claim that scores illuminate the tested topic?

Standardization of testing conditions is a substantial challenge given the greatly varying contexts and populations of American public schools. For example, to test students with disabilities, as required under NCLB, varied accommodations are necessary and required by federal law. In addition, students with markedly different degrees of English language fluency are tested. They are also given varying accommodations, depending in part on state law. Approximately15 to 20 percent of all students fall into one or both of these categories (Hoffman, 2003, Table 10), and testing them is intended to

make schools attend to their needs. However, without standardization it is exceedingly difficult to draw reasonable inferences from their scores about their learning or their needs (Heubert & Hauser, 1999; McDonnell, McLaughlin, & Morison, 1997; Pellegrino, Chudowsky, & Glaser, 2001). Because such scores are inadequately informative, they are also a very problematic basis on which to base consequences.

A second deficiency in NCLB's informative powers stems from the wide variations in procedures used to prepare students for testing. The *Standards for Educational and Psychological Testing* (the "Joint Standards"), call for examinees to have an "equal opportunity to become familiar with the test format, practice materials, and so forth" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 74). "Ideally, examinees would also be afforded equal opportunity to prepare for a test" (AERA, APA, & NCME, 1999, p.75).

Yet, there is no agreement on what such equal opportunities entail (AERA, APA, NCME, 1999, p.75). Promulgating the same state standards and tests may meet a legal threshold for equal preparation in some jurisdictions (e.g., *GI Forum*). Yet, these do not actually yield equivalent opportunities to learn the tested content (as will be discussed shortly). Schools with markedly different student populations and neighborhood conditions are markedly different environments for teaching and learning (e.g., Darling-Hammond, 2000, 2004; Lee, 2006; Natriello & Pallas, 2001). Because of this, similar scores may say very different things about the quality of teaching and learning within any given school (Stake, 1998). Using scores as the sole or primary basis for categorizing or sanctioning schools is therefore deeply problematic.

A third serious flaw in standardization and thus the informative power of NCLB's testing system pertains to its well-known goal of universal proficiency. The Joint Standards hold that when a score is used to support a criterion referenced interpretation (such as "proficiency"), "the rationale and empirical basis should be clearly presented" (AERA, APA, NCME, 1999, p. 56). Yet, the meaning of proficiency and how that is established varies considerably across the states (e.g., Olson, 2005). Therefore, even if NCLB's stated goal should somehow be achieved by 2014, this information would be worthless: All students cannot possess proficiency in reading, math, and science and yet potentially lose proficiency simply by crossing state lines.[1]

B. Deficiencies in Meeting Constructive Aims

General George Patton said that if you tell people what to do but not how to do it, "they will surprise you with their ingenuity." Educators and state policymakers who are told to increase test scores and decrease score gaps have found many ways to do this. Among these are quite a few ingenious workarounds that improve scores without attendant improvements in learning. Some may even be counterproductive to learning.

Workarounds operate at all levels of the education system under NCLB. For example, NCLB requires states to help build local education agencies' capacity to produce learning. However, states typically lack the staff, resources, and/or knowledge

to do so (e.g., Center on Education Policy, 2006). In the face of massively unattainable goals and highly likely sanctions, state actors have creatively redefined "learning." They have done this by adjusting confidence intervals, subgroup sizes, and even what counts as "two years in a row" (Center on Education Policy, 2005; Sunderman, 2006). None of this ingenuity is cognitively constructive: It does not advance students' learning or improve instruction.

Workarounds also operate at the district and school level. For example, outright cheating during NCLB-required testing has been widely reported (see, e.g., Axtman, 2005). Changing the pool of test takers by retaining students before they reach tested grades also improves scores without improving learning. So does ramping up test preparation for students nearest the passing score, while diminishing attention to other students (e.g., Booher-Jennings, 2005).

Perhaps the most common workaround in a high-stakes system is narrowing of the curriculum. About 25 percent of schools report reductions in social studies, science, arts and music in response to NCLB (Center on Education Policy, 2005). This increased focus on tested content and skills does raise scores. Unfortunately, these higher scores do not mean that students are mastering the tested disciplines: Score gains produced on a given state's high-stakes assessment are often not mirrored on other tests of the same content (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998). This pattern holds true under NCLB (Fuller, Gesicki, Kang, & Wright, 2006; Lee, 2006). As Dylan Wiliam (2001, p. 165) aptly remarked, "…the more precisely we specify what we want, the more likely we are to get it, but the less likely it is to mean anything."[2] Alas, there is one meaning that is reasonable to extract from narrowed instruction: If it doesn't prepare students to apply knowledge across tests, it isn't preparing students to apply knowledge to complex problems found in the world beyond school.

Although workarounds are pervasive, they are probably not evenly distributed. Logically, they will tend to be more frequently used in schools that are at greater risk of sanction (McNeil, 1988; von Zastrow, 2004). These include schools serving higher percentages of traditionally lower-scoring students who are supposed to be NCLB's key beneficiaries (NCLB, Sec. 1001(2)(3)(4)(5)(6)). They also include schools that have diverse student populations, in which a single errant subgroup will sink the whole ship (Kim & Sunderman, 2005; Lee & Wong, 2004; Popham, 2004). In contrast, students in affluent, homogenous schools will continue to have better access to school and classroom practices that actually promote learning. Thus, one logical consequence of NCLB's accountability system will be that substantive opportunities to learn will continue to diverge – and may increasingly diverge – on the basis of wealth, race, and ethnicity.

Part II: Toward a Better Accountability System: Clarifying the Meaning of Multiple Measures

Business people have long urged educators to take lessons from their sector. One key business lesson that education policymakers need to grasp is that an overemphasis on the bottom line corrupts the processes above it. When this occurs, improvements in the bottom line no longer clearly signal actual profits or gains.

Therefore, a good accountability system–in education or business–necessarily employs varied and multiple indicators. In education it does so for three short and related reasons: Multiple indicators make it harder to rely on workarounds. As a result, a system of multiple indicators would encourage reasonable educational practices. In turn, assessment results would better reflect improved teaching and learning and be more informative.[3]

Specifying that an educational accountability system must be informative and cognitively constructive makes it possible to eliminate some of the current ambiguities in the meaning of multiple measures. Given these two fundamental aims, the repeated administration of the same type of test *does not* constitute an acceptable system of multiple measures. From a statistical vantage point, this practice diminishes the chances of "false negatives" and thus accuracy. Yet, the repeated use of the same or similar standardized tests tends to narrow curriculum and instruction. Thereby, it undermines the teaching of the discipline and fails to be constructive (see, e.g., McNeil & Valenzuela, 2001; Shepard, 2000). In turn, test results no longer enable inferences about actual disciplinary learning. Thus, they also fail to be informative.[4]

Reliance on a content-related test together with non-test indicators, such as attendance or graduation rates allowed under NCLB, also *does not* constitute a workable system of multiple measures. These other indicators are useful. For example, they may lessen perverse incentives to alter the pool of test takers in order to improve scores (Koretz, 2003). Even so, such a system does not prevent the warping of instructional practices and gaming associated with excessive emphasis on a single test. Thus, it fails to be constructive. Thereby, it also fails to yield accurate information about learning in the disciplines: These are domains of knowledge, skills, and understandings (e.g., mathematics, science, writing) that are used and valued in the wider culture, not just on tests or in school (see Kornhaber & Gardner, 1993).

Thus, a sound accountability system must include multiple forms of assessment not just the repeated use of one test (or similar forms of it). As outlined below, a sound accountability system must also include other kinds of indicators, both to keep assessments from having perverse consequences and to monitor other important qualities of schooling (Koretz, 2003).

Part III: Beyond Standardized Testing in School Accountability: Components of a Constructive and Informative System of Multiple Measures

In Section 1111 of NCLB, the word "same" is repeatedly used to modify "standards," "expectations," "knowledge," "accountability," and "assessment." Nevertheless, states employ different tests, different definitions of proficiency, different ways of calculating AYP, and different accommodations for students. They also provide students with widely varying opportunities to learn. It is clearly possible to devise a system that is less contradictory and more cognitively constructive and informative. The components of such a system are outlined below.

A. Standardized Tests

Standardized tests are a favored accountability tool because of their strengths: Test scores appear objective: little judgment is required to evaluate whether a "95th percentile" or "advanced level" is higher than "30th percentile" or "basic ." Tests can efficiently gather large amounts of data broadly, quickly, and relatively cheaply on a per pupil basis. These data facilitate comparisons across students, schools, districts, and states. Because of its broad monitoring and reporting capacities, standardized testing tied to stakes readily enables policymakers to send and collect signals about school systems.

The strengths of standardized testing have tended to overshadow their weaknesses. Though they can probe more deeply, they tend to emphasize disjointed facts and rote skills much more so than the rich problem solving used in the disciplines (Pellegrino, Chudowsky, & Glaser, 2001). Standardized testing typically employs artificially short time frames and contexts that lack the many resources that are used to develop good disciplinary work (e.g., other people, reference materials, tools). Given these differences, an extended chain of argument, evidence, and inferences is needed to say that test results are valid representations of actual levels of disciplinary performance (e.g., proficiency in "writing" or "probabilistic reasoning"). In addition, it typically takes weeks or months before results from state accountability tests reach schools. This attenuates their usefulness for instructional purposes (Pellegrino, Chudowsky, & Glaser, 2001). In essence, standardized tests are relatively better at serving the informative, rather than the constructive, aims of an educational accountability system.

Rebalancing the position of standardized tests

There are at least two possibilities for maximizing standardized tests' strengths and minimizing their weaknesses within a system of multiple indicators. The first possibility is that the blanket, testing of all students ("census testing") should be markedly reduced (Pellegrino, Chudowsky, & Glaser, 2001). Census testing makes little economic sense: Testing is by nature a sampling enterprise. Therefore, it is costly and inefficient to generate system-wide information by testing every student each year. Census, high-stakes testing also has corrosive effects on educational practice. Thus, reducing such testing and assigning test data a minority role within an accountability system can benefit instruction and thereby improve the system's informative function.

Politically, such an approach to high-stakes testing might be a hard sell. It would raise concerns that commitments to standards, low-achievers, and disaggregated data were being abandoned.[5] Furthermore, sampling schemes would likely need to vary between, for example, large heterogeneous schools and small rural ones. The selection of students for any given sampling scheme could increase opportunities for gaming. In addition, using one test for measuring more than one purpose (e.g., growth and achievement) may increase that test's importance and the temptation to employ workarounds.

A second, and admittedly outside-the-box possibility, would be to adopt many different kinds of standardized tests and administer them all in a census fashion. (Gasp!)

However, this second possibility might be preferred by states that are committed to the rhetoric of NCLB. In fact, this second approach has some advantages: If all test results together were still to play a minority role in the overall assessment of a school or district and had to be used alongside other indicators, the futility of teaching to the tests might override the impulse to do so. This approach might avert some gaming problems associated with sampling. In a more positive vein, each test could shed light on a different indicator. For example, one test could monitor criterion-referenced achievement on the state standards; another, criterion-referenced pegged elsewhere;[6] another growth; another national norms, etc. Under NCLB, the latter two are increasingly opaque.[7]

How would either of these two possibilities enable policymakers to signal that they are serious about improving the learning of subgroups and reducing achievement gaps? First, the elimination of a single high-stakes test should reduce the excessive test preparation now visited disproportionately upon traditionally low-scoring groups. It might also free up resources that are being unduly spent on test prep materials. It could enable more teaching of richer disciplinary content. Second, using test data to monitor growth, not just status, makes learning across the full range of students more likely to matter.[8] This increases the possibility that more students' learning needs will be addressed. In contrast, under the current NCLB accountability system, it makes the most sense to attend to those students whose improvement is likeliest to put the school over the bar (see, e.g., Booher-Jennings, 2005). Under NCLB, it also makes sense to retain students in untested grades, even though retention increases the chance of dropping out (see, e.g., Hauser, 2001; Holmes, 1989).[9] Third, by eliminating excessive test preparation for a single, "same test," the accountability system will be more likely to send signals about what matters (e.g., achievement, growth, real disciplinary knowledge and skills). It should also receive more accurate signals about learning and teaching.

At the end of the day, it is fruitless for a school or district to pursue equal and gap free scores on the same test. This pursuit corrupts instruction, especially for students who have been traditionally poorly served. Instead, our education system should strive to produce cognitive equity (Kornhaber, 1998). That is, it should enable people, regardless of their ascriptive characteristics, to understand, use, and contribute to an array of disciplinary knowledge in the wider world. An accountability scheme that employs a single high-stakes test to assess every student, school, and district cannot do this. Neither can an accountability system that relies exclusively on standardized tests. The focus of such systems invariably reverts from the target (disciplinary competence) to the indicators (test scores).

B. Classroom-Based Assessment/Formative Assessments.

As important as standardized tests have become in public schools, classroom assessment remains a mainstay of educational practice. Classroom assessment ranges widely from teacher-made tests, to spot checks of cell illustrations or student graphs, to the evaluation of each step in the production of a polished essay. Such assessments guide teachers in their ongoing instruction of individual students, groups, and the entire class.

Classroom assessment can mirror the types of activities and products actually used in the practice of the discipline outside of school and engage resources commonly used to do such work. (Such assessments are sometimes called "authentic."). Relative to standardized test regimes, this provides more direct evidence of domain- relevant competence (Wiliam, 2001; Shepard, 2000). Thus, a smaller inferential leap is needed to claim that a given assessed performance represents a given, degree of disciplinary knowledge, skills, and understanding. Furthermore classroom assessment can improve actual disciplinary learning, as will be detailed shortly (Black & Wiliam, 1998a, 1998b; Stiggins, 2001; Wiggins, 1998).

Just as standardized testing does not in practice fulfill its theoretical promise, neither does classroom assessment. However, unlike standardized testing, classroom assessment's problems have overshadowed its potential strengths. For example, classroom assessment, reinforced by high-stakes testing, tends to emphasize rote knowledge and skills (Black & Wiliam, 1998; Shepard, 2000). Though classroom assessment has been used in large-scale accountability systems (e.g., Bandalos, 2004; Koretz, 1998), it is cumbersome to do so. Such assessment must also draw on informed judgment. Judgments made by different raters of student work have often been inconsistent (e.g., Koretz, 1998; Linn & Baker, 1996).

All told, compared to standardized tests, classroom assessment advances cognitively constructive aims relatively better than informative aims. Nevertheless, classroom assessment must be developed and valued within an accountability system (See Stiggins, 2001; Pellegrino, Chudowsky, & Glaser, 2001). If it is not, workarounds and gaming will infiltrate educational practice. As a result neither the informative nor the constructive aims of an accountability system will be met, which is now the case under NCLB.

Rebalancing the position of classroom assessment

Building teacher capacity: The potential strengths of classroom assessment lie in the development of teachers' capacity to use it formatively (Shepard, 2000; Stiggins, 2001, Black & Wiliam, 1998a). Formative assessment entails a clear understanding of what good work is (the desired goal), the ability to help students to develop this understanding, and the capacity to give students specific, discipline-relevant feedback, which fosters good work (Black & Wiliam, 1998a). Teachers can use formative assessment in reviewing and grading exams. For example, a teacher can point out that although students' answers on an algebra exam were correct, their solutions relied more on arithmetic- rather than algebraic problem solving. The teacher can then point out the difference between these approaches and work with students in class to develop algebraic approaches. Teachers can also use formative assessment during instructional time, for example, when they highlight for the class how a given student's essay has incorporated descriptive and figurative language that are valued in the discipline (see, e.g., Kornhaber, Fierros & Veenema, 2004, Chapter 6).

When formative assessment is carried out, student investment and their learning increase (Black & Wiliam, 1998a, 1998b; Chappius & Stiggins, 2002, Crooks, 1997,

Wiliam, 2001, 2004). In a meta-analysis of over 200 studies, formative assessment was found to produce overall gains in learning, and gains that were bigger than existing educational interventions for students with low achievement levels (Black & Wiliam, 1998a, 1998b).

Formative assessment nevertheless raises two feasibility questions. First, "Can teachers really 'do' formative assessment?" In fact, there is evidence that teachers below the level of Hollywood's Mr. Holland or Jaime Escalante can deploy formative assessment competently. For example, Wiliam (2004) has found that teachers can learn formative assessment practices, given interest and brief opportunities to discuss their efforts. A non-generalizable but intriguing finding is that students of such teachers performed markedly better on an external exam relative to a comparison group whose teachers did not focus on formative assessment (Wiliam 2004, 2006). Furthermore, formative assessment has long been used by many ordinary teachers, including coaches (Wiggins, 1998) and teachers of music and other arts disciplines. Relatedly, student work in these areas more frequently mirrors the good disciplinary work carried out beyond school walls.

None of this is meant to sugar coat the work that would be entailed in widespread development of formative assessment in regular classrooms. Teachers generally like their tests to mirror standardized tests (Shepard, 2000). This preference is reinforced by the prevalence of high-stakes standardized tests (McNeil, 1988; Shepard, 2000). Enacting good classroom assessment is also stressful to teachers (Bandalos, 2004, Stecher, 1998). Another difficulty is that classroom teachers are ill-trained to use formative assessment (Shepard, 2000; Stake, 1998; Stiggins, 2001). Few states require that teachers or administrators have assessment competence, even though teachers may spend as much as one-third of their time carrying out a range of assessment activities (Stiggins, 2001). Developing teacher know-how through changes to teacher certification and professional development efforts could gradually address this difficulty (Shepard, 2000; Stiggins, 2001). In essence, formative assessment does not provide any instant answers to meeting the informative and constructive aims of accountability systems. However, if the role of standardized testing is markedly reduced in such systems, the difficulties associated with developing formative assessment practices seem ultimately tractable to policy solutions.

Creating public audiences. The second feasibility question concerning formative assessment is "Is it technically "good enough?" For example, ratings of portfolios of student work vary widely and for multiple reasons (Koretz, 1998; see also Baker & Linn, 1996). This lack of consistency makes it is difficult to support inferences about the quality of students' performances. Holistic scoring (i.e., an overall rating, rather than ratings based on different dimensions of the work) is more consistent (Koretz, 1998, Pellegrino, Chudowsky, & Glaser, 2001). However, the potential sources of this consistency may still make it difficult to draw appropriate inferences. For example, given a collapsed range of scoring, chance alone would increase consistency in ratings because there are fewer possible scores to assign (Koretz, 1998). Furthermore, efforts to achieve scoring consistency by reducing criteria or score ranges could erode teachers'

and students' consideration of the multiple processes and dimensions of good disciplinary work, which is central to good formative assessment.

One way out of this box is to consider that across the disciplines in which formative assessment is widely used in school (e.g., the arts, sports), there is a public audience for the work. Correspondingly, the absence of spectators for most of academic endeavors in classrooms could well allow low or idiosyncratic standards to dominate formative assessment. One way to avoid this threat is to form consortia of schools or small districts to develop formative assessment practices. This has been done in Nebraska's STARS (Student-based Teacher-led Assessment and Reporting System) system, which gives a central role to formative assessments developed at the district level by teachers (Bandalos, 2004). Consortia members consisted of small districts that worked together to devise assessment plans that entailed developing curriculum and assessments aligned to state standards. Some consortia hired experts to help them develop this work (Bandalos, 2004).

Another approach to creating an audience is to institute a system of external evaluators. Britain's former system of school inspectors provided in-depth reports that were seen as quite valid representations of a school's particular strengths and needs, even though such reports were not produced in a consistent fashion (Smith, 2000). The inspectorate was comprised of highly regarded former teachers and administrators who performed something like critical friends (Bolton, 1998). A former head inspector has argued that such an external system of reviewers might operate with agreed upon guidelines, but must still have substantial flexibility to exercise judgment (Bolton, 1998).[10] Mirroring classroom formative assessments, inspectors' reports could provide much more specific and fine-grained information to schools and the public than is now provided by high stakes tests.

An audience might also be built by networking groups of schools involved in formative assessment to one or more colleges or universities. Such institutions, at least in theory, maintain a mission of service as well as a concern for teaching and producing disciplinary expertise and new knowledge. Consortia that networked clusters of schools and colleges could also help address other problems. They could improve the articulation of standards between K-12 and higher education. In turn, this could help fend off increasing calls for test-based accountability at the college level and the downward spiral in assessment (and learning) that would likely ensue.

It will take considerable effort to employ formative assessment practices. However, teachers can acquire these skills. Audiences can be developed and thereby reduce the risks of schools relying on idiosyncratic or low standards. Policies can be devised to address these problems. Given this, formative assessment ought to be able to play another minority role in an accountability system.

C. Other Indicators

Balancing formative and high-stakes assessments make it more likely that the informative and constructive aims of an accountability system could be met. However,

as Stake (1998) has noted, "Schooling includes many performances, provisions, and relationships which could be assessed…" Including some of these in an accountability system is important to the overall functioning of teaching, learning, and the system itself.

Inputs

Clearly, there are many inputs that would foster more equal achievement of students in schools. In this paper, I am focusing only on inputs related to developing a constructive and informative system of multiple indicators.

States should require that anyone certified to teach must have substantial coursework in formative assessment. This would in turn require colleges of education to devise such courses. Stiggins (2001) has noted that content area specialists in colleges of education could work together with measurement specialists to develop formative assessment strategies for diverse curricular areas.

States should also be required to engage in public education efforts to help parents and the wider community to understand the strengths and weaknesses of different forms of assessment. This information could be made available on state education department websites and also distributed within schools. This information can help to create a public space for the development of formative assessment practices.

Processes

Professional development is an essential component of constructive, and therefore, informative accountability systems. Schools, districts, and states therefore ought to be evaluated on the quality and accessibility of professional development aimed at enhancing formative assessment practices in classrooms.

Outputs:

Retention rates: Since retention is strongly associated with drop, an accountability system should require schools and districts to report retention rates and efforts to minimize retention.

Graduation rates: States should be required to put into place systems that accurately collect and report disaggregated data about students who are graduating with a standard high school diploma in four years (Losen, 2004; Swanson, 2004). Districts and high schools should be held accountable for accurate reporting within this system and for efforts to increase standard, four-year graduation rates. Ideally, schools and districts should also produce evidence that their graduates are moving into higher education, training, or jobs.

Compensatory combinations

NCLB requires the conjunctive use of test scores, high school graduation rates and at least one other indicator of the state's choosing to determine whether schools and districts are succeeding. That is, adequate performance on all three indicators must be in place to avoid sanction. In a system of multiple indicators, it would be more useful if standardized tests, formative assessment, and other indicators could be used in a

compensatory fashion, with states setting out reasonable parameters for performance within each area. That is, somewhat higher performances in some areas could offset somewhat lower performance in other areas. Schools and districts could then work within these parameters depending on their local situations. For example, some communities and their teachers may want to emphasize improving graduation rates and formative assessment practices, networks and results, more than improving test scores. In other communities, it may make sense to give more equal weight to test scores, formative assessment practices and results, and other indicators. This compensatory approach provides a degree of flexibility that NCLB has promised, but not delivered.

IV: Some suggested changes to NCLB

Enacting many of the ideas discussed to this point may have to wait for a post-NCLB world. However, here are a few possible areas of the legislation that may be tackled in the interim. Doing so may help to develop an accountability system that is truly informative and cognitively constructive.

1. NCLB calls for the "same accountability system" (Sec. 1111(b)(2)(A)(ii)) to be used within states. However, any potential for such a system to include formative assessment is undermined by Section 1111(b)(3)(C), which requires assessments that are consistent with professional and technical standards of reliability and validity to serve as "the primary means of determining yearly performance of the State." Given the problems with exclusive reliance on any single type of assessment, this should be modified: Assessments whose use is consistent with professional and technical standards for reliability and validity should be one, but not necessarily the primary, means of assessing schools and districts. (See Part III, above.)

Despite Nebraska's struggle with the Department of Education (Christensen, 2006, Johnson, 2006), NCLB could be brought to heel by revisiting the language of "same accountability system." Ideally, the definition of "same" should be that both tests that meet psychometric standards as well as publicly-monitored formative assessment can or should be included in the determination of school, district, and state adequacy.

2. NCLB Section 1111(b)(2)(C)(vii) allows states to include at least one other indicator besides high school graduation rates, e.g., retention, attendance, and students in advanced placement courses. However, 1111(2)(D) requires those indicators are employed in ways that are consistent with professional and technical standards for validity and reliability. Further, Section 1111(2)(D)(ii) does not allow additional indicators to modify the categorization of a school as needing improvement, corrective action or restructuring. This means that additional indicators serve as window dressing: the substance of accountability remains essentially focused standardized test scores.

These paragraphs need to be modified to *require* other indicators to be included (e.g., those mentioned in Part III of this paper). Some parameter-driven compensatory combination of these should be allowed to count substantively toward improvement of a school's categorization; otherwise, little attention will be given to improving other

important aspects of schooling. In addition, including some of these other indicators may help to limit workarounds (see Koretz, 2003).

3. Section 1111(b)(2)(A)(iii) requires sanctions and rewards to hold schools and LEAs accountable for student achievement and making adequate yearly progress. This should be modified to include not just sanctions and punishments primarily on the basis of test results, but also on the basis of incorporating some of the other indicators.

4. NCLB Section 1111(b)(2)(C)(v) requires disaggregation of data except when subgroups are too small to be statistically reliable or will lead to the identification of individual students. In the latter case, subgroup scores should not be released. In the former, information about subgroup performance should nevertheless be made public to encourage attention to these students' learning.

5. Finally, sections of NCLB that address the provision of highly qualified teachers should require states to set in motion the development of certification and professional development programs that enable teachers to acquire and use formative assessment competently within the classroom.

V. Conclusion

With considerable effort, policy makers can foster a system of assessment and accountability that can diminish the corruption of learning and information that currently exist under No Child Left Behind. However, no accountability system can do what NCLB has promised – create universal proficiency as measured against high standards. No society has ever achieved that. Furthermore, variability among individual students has multiple sources. The variability in school achievement across groups is spurred primarily by processes, policies, and practices beyond K-12 education. However, education can still benefit from a good accountability system. A good system would not wave unrealistic banners. Instead, it would strive to meet two criteria: to build cognitively constructive educational practices and to generate accurate information about the status and development of student learning.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). IStandards for educational and psychological testing. Washington, DC: AERA, APA, NCME.

Axtman, K. (2005, January 11). When tests' cheaters are the teachers. *Christian Science Monitor, online:* http://www.csmonitor.com/2005/0111/p01s03-ussc.htm

Balou, D., Sanders, W. & Wright, P. (2004).. Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Bandalos, M. M. (2004). Can a teacher-led state assessment system work? *Educational Measurement, Issues and Practice, 23*(2), 33-40.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice, 5*(1), 7-74.

Black, P., and Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Bolton, E. (1998). HMI—The Thatcher years. *Oxford Review of Education, 24*(1), 45-55.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Buckendahl, CW, Plake, B.S., & Impara, J.C. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement, Issues and Practice, (23*(2), 14-26.

Center on Education Policy (2005). NCLB policy brief 3: Is NCLB narrowing the curriculum? Washington, DC: Center on Education Policy. Available at: http://www.cep-dc.org/nclb/NCLBPolicyBriefs2005/CEPPB3web.pdf

Center on Education Policy. (2006). *From the Capitol to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center on Education Policy Available at: http://www.ctredpol.org.

Chappius, S., & Stiggins, R.J. (2002). Classroom assessment for learning. *Educational leadership, 60*, 40-43.

Christensen, D. (2006, July 5). Statement of the Commissioner of Education regarding state's assessment system and its compliance with the requirements of NCLB.

Crooks, T.J. (1988). The impact of classroom evaluation on students. *Review of Educational Research, 58*(4), 438-81.

Darling-Hammond, L. (2000). New standards and old inequalities: School reform and the education of African American students. *The Journal of Negro Education, 69*(4), 263-287.

Darling-Hammond, L. (2004). Inequality and the right to learn: Access to qualified teachers in California's public schools. *Teachers College Record, 106*(10), 1936-1966.

Fredericksen, J.R, & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006) Is the No Child Left Behind Act working? The reliability of how states track achievement. Working Paper 06-1 University of California, Berkeley: Policy Analysis for California Education. Available at http://pace.berkeley.edu/testscoretrends.html.

Gardner, H. (1999). *The disciplined mind.* New York: Penguin/Putnam.

GI Forum, Image de Tejas v. *Texas Education Agency,* 87 F. Supp. 667 (W.D. Tex. 2000).

Hauser, R.M. (2001). Should we end social promotion? Truth and consequences. In G. Orfield & M.L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*, pp.151-178. New York: Century Foundation Press.

Heubert, J., & Hauser, R.M. (Eds.). (1999) *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

*Hoffman, L.M. (2003), NCES/CCD tables

Holmes, C.T. (1989). Grade level retention effects: A meta-analysis of research studies. In L. Shepard & M.L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16-33). London: Falmer Press.

Johnson, H. L. (2006, June 30). Nebraska assessment letter. Available at: http://www.ed.gov/admins/lead/account/nclbfinalassess/ne2.html

Johnson, H. L. (2006, August 8). Decision letter on request to amend Nebraska accountability plan. Available at: http://www.ed.gov/admins/lead/account/letters/acne6.html

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., & Stecher, B.M. (2000). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koretz, D. (1998). Large scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, policy & practice, 5*(3), 309-334.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice,* 22(2), 18-26.

Koretz, D. & Barron, S. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

Kornhaber, M. L. (1998). A roundtable: The black-white test score gap. *The American Propsect, 41,* 64, 66.

Kornhaber, M. L. & Gardner, H. (1993). Varieties of excellence: Identifying and assessing children's talents. New York: The National Center for Restructuring Education, Schools, and Teaching (NCREST), Teachers College, Columbia University.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. American Educational Research Journal, 41(4), 797-832.

Lee, J. (2006). Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcomes trends. Available at: http://www.civilrightsproject.harvard.edu/news/pressreleases/nclb_report06.php

Linn, R.L. (2004) Rethinking the No Child Left Behind accountability system. Paper prepared for a forum on No Child Left Behind sponsored by the Center on Education Policy. Washington, D.C., July 28, 2004.

Linn, R. L. & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth Yearbook of the National Society for the Study of Education, Part I* (pp. 84-103). Chicago: University of Chicago Press

Losen, D. J. (2004). Graduate rate accountability under the No Child Left Behind Act and the disparate impact on students of color. In G. Orfield (Ed.), *Dropouts in America: Confronting the graduation rate crisis*. Cambridge, MA: Harvard Education Publishing Group.

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*, pp. 85-106. New York: Century Foundation Press.

McDonnell, L. M., McLaughlin, M.J., & Morison, P., (1997). *Educating one and all: Students with disabilities and standards-based reform.* Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

McNeil, L. M. (1988). *Contradictions of control: School structure and school knowledge.* New York: Routledge.

McNeil, L. M., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*, 127-150. New York: Century Foundation Press.

Natriello, G., & Pallas, A.M. (2001). The development and impact of high-stakes testing. In G. Orfield & M.L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*, pp. 19-38. New York: Century Foundation Press.

No Child Left Behind Act of 2001 Pub. Law No. 107.110

Olson, L. (2005). Room to maneuver. A progress report on the No Child Left Behind Act, *Education Week*, pp. S1-S6, December 14.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). Knowing what students know. Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

Popham, W.J. (2004). Ruminations regarding NCLB's most malignant provision: Adequate yearly progress. Paper prepared for a forum on No Child Left Behind sponsored by the Center on Education Policy. Washington, D.C., July 28, 2004.

Sanders, W., & Horn, S.P. (1998). Research findings from the Tennessee value-added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247-256.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Singley, M. K., & Anderson, J. R. (1989). *; Transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Smith, G. (2000). Research and inspection: HMI and Ofsted, 1981-1996: A commentary. *Oxford Review of Education, 26*(3-4) 333-352.

Stake, R. (1998). Some comments on assessment in U.S. education. *Education Policy Analysis Archives, 6*(14). Available at: http://epaa.asu.edu/epaa/v6n14.html

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education: Principles, Policy & Practice, 5*(3)

Stiggins, R.J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement, Issues and Practice, 20*(3), 5-15.

Suen, H.K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review, 58* (1), 46-65.

Sunderman, G., (2006) The unraveling of No Child Left Behind: How negotiated changes transform the law.

Swanson, C. B. (2004). Sketching a portrait of public high school graduation: Who graduates? Who doesn't? In G. Orfield (Ed.), *Dropouts in America: Confronting the graduation rate crisis*. Cambridge, MA: Harvard Education Publishing Group.

von Zastrow, C. (2004). Academic atrophy: The condition of the liberal arts in America's public schools. Washington, D.C.: Council for Basic Education. Available at: http://www.ecs.org/html/offsite.asp?document=http%3A%2F%2Fdownloads%2Encss%2Eorg%2Flegislative%2FAcademicAtrophy%2EpdfConclusion:

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass.

Wiliam, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.), *Curriculum and assessment*, pp. 165-181. Greenwich, CT: JAI Press.

Wiliam, D. (2004). Keeping learning on track: Integrating assessment with instruction. Invited address to the 30th annual conference of the International Association for Educational Assessment. Philadelphia, PA. June 2004.

Wolf, D.P., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well. *Review of Educational Research, 17*, 31-74. Washington, DC: American Educational Research Association.

Endnotes

[1] It is worth noting that the problem of defining and developing students' proficiency would not be eliminated by a single national standardized test. Culturally valued disciplines, such as writing, mathematics, and science are too multifaceted to be adequately evaluated by a single measure (Gardner, 1999; Pellegrino, Chudowsky, & Glaser, 2001; Wiggins, 1998; Wolfe, Bixby, Glenn & Gardner, 1991). Therefore, meeting a standard of "proficient" on one test cannot indicate proficiency in the wider domain. Furthermore a single test would exacerbate gaming to improve the score. This undermines the ability to infer increased scores indicated increased learning.

[2] Transfer of knowledge and skills from one context to another has long been a central problem in schools (Singley & Anderson, 1989). However, NCLB's excessive emphasis on attaining score gains and sanctions has not improved this problem (Lee, 2006) and is likely only to exacerbate it (Kornhaber, 2004).

[3] There is also an essay answer: A few civilizations have been known for their testing schemes (Suen & Yu, 2006), and ours may now be among them. However, test scores do not capture any civilization's genuine accomplishments (Gardner, personal communication, April 2005). Civilizations, and individuals within them, are instead recognized for their contributions to literature, mathematics, systems of government, philosophy, technology, science, architecture, art, and other disciplines (Gardner, personal communication April 2005). With multiple indicators, educators are more likely to expose students to the multifaceted knowledge, skills, and methods of analysis of those disciplines that are taught in school. In turn, this may engage more students and foster their understanding of, and capacity to contribute to, those disciplines.

[4] Relatedly, the use by states and districts of multiple preliminary pretests may provide information needed to boost pass rates, but it does not alleviate the problems of narrowing of curriculum and lack of transfer. This practice is a rational, but not educationally sound, response to systemic overreliance on a single high-stakes test.

[5] Under NCLB, the latter is being abandoned as subgroup sizes and confidence intervals are allowed to increase by the Department of Education.

[6] NAEP does this at the state level, but it can't be used for individual schools or for most school districts.

[7] It is possible to imagine using multiple tests in a sampling fashion. In practice, the complications around sampling and gaming would probably swamp such a system.

[8] Various approaches exist to detect and convey growth in student learning. Value-added assessments have attempted to show the extent to which schools and teachers have contributed to student growth (Sanders & Horn, 1998). Although these models have been criticized for not fully parsing out background and sociological influences (Balou, Sanders & Wright, 2004), they are increasingly being adopted. Popham (2004) has argued that, despite their weaknesses, grade level equivalent tests might be employed to

see whether a year's growth is occurring in a year's time. This approach is relatively transparent. However, it does not typically rest on any knowledge of how students' conceptual knowledge and skills actually develop within disciplines. Ideally, assessments should be built on "progress maps." These "attempt to capture in words and examples what it means to make progress or to improve in an area of learning" (Pellegrino, Chudowsky, & Glaser, 2001, p. 137). Assessment based on progress maps could provide "teachers, parents, and administrators with a shared understanding of the nature of development across the years of school and a basis for monitoring individual progress from year to year" (Pellegrino, Chudowsky, & Glaser, 2001, p. 137). Even so, a progress-map measurement should still play a minority role in school accountability and decisions for individual students. Historical and cross-cultural evidence demonstrate that no matter what its form or the intention behind it, any test that is overused and has serious consequences will contort teaching and learning (Suen & Yu, 2006, Madaus & Clarke, 2001).

[9] NCLB and later revisions to it also encourage state level policymakers to make whole subgroups of students disappear by enlarging the subgroup sizes. One way to reduce this tendency is to publish the average scores of all groups, but indicate that some groups are too small to allow inferences from the score to be used for specified purposes.

[10] Bolton (1998) has also asserted that such inspectors should not be under the control of policymaking bodies. This will enable their reports to be more aligned with educational concerns than the expediencies of policymakers (Bolton, 1998; Smith, 2000).