Center for
K–12 Assessment
& Performance Management

*An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.*

**Exploratory Seminar:**

Measurement Challenges Within the Race to the Top Agenda

December 2009

# Issues in Measuring Student Growth and Conducting Productivity Analyses

Henry Braun

*Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.*

# Issues in Measuring Student Growth and Conducting Productivity Analyses

Henry Braun

Boston College

This paper was presented by Henry Braun at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of other papers presented at the seminar at http://www.k12center.org/publications.html.

The Obama administration has proposed making an unprecedented investment in educational innovation through its Race to the Top initiative. Funds allocated to this initiative amount to $4.35 billion, which includes $350 million,

> …to support consortia of States working toward jointly developing and implementing a next generation of common summative assessments that are aligned with a common set of K–12 internationally benchmarked, college and career ready standards that model and support effective teaching and student learning. (U.S. Department of Education, 2009, p. 1)

The U.S. Department of Education provides a framework of desiderata that is supplemented by a set of 10 required and 4 desired characteristics. The desiderata are that the summative assessments should measure:

- Individual student achievement as measured against standards that build toward college and career readiness by the time of high school completion;
- Individual student growth (that is, the change in student achievement data for an individual student between two or more points in time); and
- The extent to which each individual student is on track, at each grade level tested, toward college or career readiness by the time of high school completion. (U.S. Department of Education, 2009, p. 2)

A close reading of these desiderata reveals a number of implicit assumptions, among them that one can sensibly talk of college and career readiness for the general population, that it is possible to build an articulated set of academic standards that are predictive of future accomplishments, and that it is meaningful to measure student growth through performance on a series of academic assessments. More critically, one can ask whether a new system of summative assessments, however radical and innovative, can spearhead rapid and substantial educational progress. Each of these assumptions deserves extensive and thoughtful treatment. To add to the challenge, the Department of Education expects that, in addition, these assessments will be useful not only in the determination of the

effectiveness of schools, principals, and teachers, but also provide information to support the improvement of teaching and learning.

To say this is an ambitious undertaking would be a gross understatement. The technical, logistical, and political obstacles are substantial. Nonetheless, as Education Secretary Duncan has reminded us, the Race to the Top program is an unprecedented opportunity for the education community to make fundamental breakthroughs and to make tangible progress in providing a quality education for all the nation's students. It behooves us, then, to do our best in providing thoughtful commentary on the agenda of the Department of Education, as well as concrete suggestions on how to move forward.

In this paper, I will focus on the twin issues of assessment design and accountability, especially as they relate to so-called productivity analyses of educators and education systems. Given the breadth and complexity of the topic, I will only be able to touch on a few salient aspects, with the hope that the treatment will inform both policy discussions and the planning of the consortia. The paper begins with a brief overview of the current policy landscape, which is followed by sections on assessment and accountability. It concludes with a discussion of some salient issues related to this initiative.

## Education Policy Landscape

At the federal level, there is palpable impatience with the slow progress in raising achievement for all students and narrowing the achievement gaps that have remained persistently large over the last two decades (Braun, Chapman, & Vezzu, 2010; Braun, Wang, Jenkins, & Weinbaum, 2006; Lee, 2007; Lee, Grigg, & Dion, 2007). These problems are compounded by high dropout rates, especially among disadvantaged and minority students, and incontrovertible evidence that many high school graduates are not prepared with the basic cognitive (and other) skills needed for life after secondary school (Education Week, 2009). For example, it is not uncommon for community colleges to find that upwards of 40% of recent high school graduates require developmental courses (i.e., remediation) in mathematics, reading, or both, before they can register for credit-bearing courses.

It is clear to most observers that the No Child Left Behind Act of 2001 (NCLB) has, in general, not succeeded in bending the achievement curve and, indeed, has caused much collateral damage (Koretz, 2008; Madaus, Russell, & Higgins, 2009). The Obama administration is now planning the reauthorization of the Elementary and Secondary Education Act (ESEA) and is grappling with how to reconcile the desire for meaningful and constructive accountability with the limitations imposed by the need to respect states' autonomy, the current status of educational measurement, and the tools for evaluating educational effectiveness, as well as the truism encapsulated in Campbell's Law (Campbell, 1979):

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (p. 31)

Though this is not quite a law of nature, it does concisely summarize empirical experience over a broad range of fields (see also Rothstein, 2008). The lesson, then, is not that one should dismiss or ignore the law; rather, one should try to design accountability systems to mitigate its predicted negative effects.

At the state level, in response to the mandates of NCLB, states have devised, among other things, content standards and assessment systems in reading and mathematics for grades 3–8, complemented by achievement standards against which to judge student performance. More recently, science assessments for selected grades have come on line. The assessments vary widely across states, both in their degree of alignment to the content standards and in their psychometric quality. Similarly, the performance standards vary in quality and rigor and are generally poorly articulated across grades. Although nearly all states have signed on to the common core standards initiative spearheaded by the National Governors Association (NGA), Council of Chief State School Officers (CCSSO), and Achieve, Inc., the states' degree of commitment to adopting the standards and developing aligned assessments is questionable. Their hesitation appears to be a function of political considerations (e.g., resentment of continuing federal intrusion in education), concerns about the financial burdens involved (notwithstanding the promised federal funding), and the lack of capacity in many state departments of education to design and implement a new system of assessments or even to properly supervise the work of outside contractors.

At present, most states are preparing applications for the Race to the Top program and the associated competition for developing next-generation summative assessment systems. It remains to be seen how the parameters of the competition and the promised funding play out in the current context of state policies, capacities, and constraints. What is not in doubt is that the next generation of assessments, and the accountability systems in which they will be embedded, will profoundly affect the nature of schools and schooling for years to come.

# Assessment Design

## Preliminary Considerations

Assessment design should not be conducted *in vacuo*; rather, it ought to be shaped not only by domain-specific considerations and generic psychometric requirements, but also by the purposes to be served by the assessment results. This is a fundamental premise of evidence-centered design (Mislevy, Steinberg, & Almond, 2002). That is, one should ask what are the inferences or decisions that will be made on the basis of the results, what evidence is needed to support these uses and, finally, which questions or probes (in what combinations) will generate such evidence? In this instance, in addition to the usual demands, there is a requirement that the assessments provide evidence with respect to student growth in the domain. Before such a design effort can commence, a number of prerequisites must be in place:

- A comprehensive model of the domain that includes the requisite knowledge and skills,
- Models of student learning in the domain that represent the principal pathways to the development of expertise,
- Content standards that are vertically articulated across grades, and

- Performance standards that are appropriately linked to the content standards and are also vertically articulated across grades.

It should be emphasized that these are prerequisites not just for the proposed assessment system but for a comprehensive and coordinated educational system that would encompass curriculum and instruction, as well as assessment. In particular, the emphasis on vertical articulation is a response to the inefficiency and misinformation that ensue when each grade, and especially each grade segment (i.e., elementary school, middle school, high school) is treated *sui generis*, with little regard for what came before and what is to come afterward.

Although a fully fleshed out exemplar does not exist, there is substantial source material available to inform the establishment of some of these prerequisites (Donovan, Bransford, &Pellegrino, 1999; Pellegrino, Chudowsky, & Glaser, 2001), particularly the cognitive bases for learning and assessment. Another report (Wilson & Bertenthal, 2006) provided a detailed description of many of the components in the context of science assessment. What is required is an approach that integrates cognitive and developmental perspectives in concert with more traditional psychometric and logistical considerations. There has also been considerable progress in the production of guidelines for developing performance standards (Hambleton & Pitoniak, 2006). Establishing appropriate and instructionally useful content standards is not a trivial task, as is apparent from the criticisms that have been directed at the standards currently in place in many states. The challenge is all the greater when the aim is to develop national rather than state standards. For an incisive review of the issues and some thoughtful suggestions, see Barton (2009).

Messick (1994) and others have argued that assessment design should be construct-based rather than task-based. That is, the delineation and clarification of the target constructs should precede task development, particularly when the tasks will be employed in periodic assessments (e.g., annually), with moderate- to high-stakes attached to the results. A construct-based approach offers a stronger basis for subsequent validity analyses and, collaterally, a better foundation for the generation of tasks that are substantively and psychometrically equivalent. As Wiley (2002) reminded us, though, the choice of constructs is deserving of its own validity analysis.

Although we now have considerable experience in working out the central constructs of a domain for purposes of instruction and assessment, we have less experience with conceptualizing the construct of growth in a domain. Although it seems straightforward enough, the problem grows in subtlety the more one thinks about it. Unfortunately, it is too easy to be misled by analogies to the measurement of growth with respect to physical characteristics like height and weight. In such cases we have scales of measurement that are interval scales; that is, a change of one unit has the same meaning anywhere along the scale. Thus, it makes sense to calculate the difference of measurements taken on one individual at two points of time and compare it to the difference taken on another individual at the same (or other) points of time. Even in this happy setting, however, the utility or value associated with a given difference may vary with the locations of the original values on the scale of measurement. For example, the gain of 10 pounds for two individuals each of height 5’ 8” with initial weights of 140 pounds and 190 pounds may signal different changes in health.

In the world of educational measurement, the question of whether test scales possess the interval scale property is subject of voluminous literature. Chapter 3 of Braun, Chudowsky, & Koenig (2010) has a discussion focused on the implications for value-added analysis. Suffice to say that the majority of psychometricians are doubtful that the scales on which the results of most large-scale standardized tests are reported possess the interval scale property. The doubt is only magnified when scores are placed on a single scale constructed through a vertical linking procedure that purportedly connects performances on different tests given at different grade levels (Briggs & Weeks, 2009; Martineau, 2006). In this case, although one can certainly calculate a difference of scores, it is not clear what that difference means and, consequently, what a comparison of score differences associated with two individuals starting at different points on the scale may signify.

Indeed, since for most tests we are hard put to offer a substantively grounded interpretation of a single score, it should come as no surprise that we are stymied when asked to interpret a difference of such scores. More problematic is the fact that most policymakers are oblivious to these difficulties, and so blithely assume that being able to calculate differences and carry out arithmetic operations on them signifies that they can be interpreted as easily as differences in height or weight. One result is that the demands placed on tests often exceed what they are able to bear and, consequently, misinterpretations abound.

At the same time, it should be acknowledged that a properly designed coherent educational assessment system could provide more useful information than do the systems in place today. A little-recognized problem is that as assessment systems improve in their alignment with respect to comprehensive content standards, there will be a concomitant increase in the complexity of the outcomes generated. Presumably, the derived growth indicators should reflect this. Current accountability systems, however, are typically not structured to deal with such complexity; the resulting mismatch must be addressed somehow.

Returning to the problem of measuring growth, the first challenge is to consider the different ways in which we might conceptualize the *growth construct*. As indicated above, a more nuanced approach to the measurement of growth presupposes both theoretical and empirical understanding of pathways to expertise. With respect to assessment design, it poses greater challenges in balancing goals and constraints (Braun, 2000). Unfortunately, we have little to guide us on the tradeoffs in obtaining improved cross-grade articulation and more interpretable measures of growth. That is, what test properties may be degraded or lost as a result?

Whatever the choice of the growth construct and the assessment instrument, the resulting growth indicator will have to be validated in different ways: its psychometric properties (e.g., reliability), its relationship to other measures of learning (a form of convergent validity), its ability to predict future performance, and its consequential validity (i.e., the direct and indirect implications of the choice for students' education).

## Further Considerations

As indicated above, the assessment systems called for by the notice for the Race to the Top (U.S. Department of Education, 2009) must serve a variety of information needs, including the support of

student learning, program/school improvement and accountability at various levels. The notice recognizes that the system will have to comprise multiple components administered throughout the academic year. This is wise, as a single time-constrained assessment cannot generate sufficient evidence to meet these varied requirements.

The design of an assessment system, rather than a single test, immediately places on the designer an obligation to regard coherence as an overarching constraint. The literature notes at least two forms of coherence: horizontal and vertical. The former refers to the situation in which such artifacts as curriculum, instruction, standards, and assessment are all grounded in a common model of cognition, learning, and representation. The latter refers to the articulation of these artifacts across grades so that transitions from one grade to the next are as seamless as possible. This is an ambitious agenda. As one report puts it: "Coherent assessment systems do not develop by accident; they must be deliberately designed" (Wilson & Bertenthal, 2006, p. 5)

What, then, might such a system look like and, specifically, how would it support various productivity analyses? Recalling the prerequisites listed above, the system would be in a dynamic relationship with other nodes in a web of educational systems, all anchored to a common foundation. In comparison to current state tests, the system should:

- Enhance construct validity through better alignment to the content standards, with respect to both breadth and depth;
- Improve systemic validity by reducing incentives to narrow the curriculum or to corrupt the testing process; and
- Lead to greater use of technology platforms.

To meet the information needs delineated in the notice, I would argue that one such system would consist of four related components:

1. *Diagnostic.* This component does *not* serve a summative role; rather, it provides instructional support, for both teachers and students, at frequencies determined by their needs. Ideally, most of these assessments would be technology-based so that the sequence of exercises or probes could be individually adapted, and feedback would be essentially instantaneous. Employing a technology platform in a low-stakes setting provides an opportunity for working out the kinks and giving stakeholders an opportunity to become familiar with the platform and its capabilities.
2. *Extended projects.* This component targets the higher order skills called for in the content standards and would offer an opportunity to challenge students with integrated tasks that could extend over a week or more. There might be two or three in the year. Some of these projects could be technology-based but would be teacher marked. For summative purposes, an audit system would review assigned grades.
3. *On-demand (extended response).* For this component, students are given prior access to source materials and then asked to provide extended responses to one or more questions based, at least in part, on the source materials. These are time-limited in-class activities, which may be

technologically based. Responses would be centrally marked. Two or three during the course of the year would be sufficient.

4. *On-demand (short response).* This component most resembles current end-of-course standardized assessments, comprising various forced-choice formats, as well as short answer responses. It would typically be administered toward the end of the school year and would be centrally marked. Over time, most states would gravitate toward using a technology platform for this component.

I believe that such a system, if properly constructed, would be consistent with the characteristics of a *high-quality assessment system* as presented in Wilson and Bertenthal (2006; p. 28). Of course, one can object to such a system on a number of grounds: it would be too time consuming, too burdensome on students and teachers, and the in-class components would introduce variations in administration that would undermine comparability. Each objection (and others not listed) has some validity and would have to be considered in the design process. It is certainly the case that an assessment system that generates valid evidence with respect to the full range of content standards is going to be big and expensive. The key is to amortize the required investment through better integration of instruction and assessment and stronger linkage of other efforts (such as teacher professional development) to the assessment process.

For most of the following discussion of productivity analysis, I will assume, for the sake of tractability, that a single test score adequately summarizes a student's current status. The final subsection will return to the question of more complex ways in which students' status can be represented and the implications for constructing measures of growth and conducting evaluations of effectiveness.

# Productivity Analysis

## Keeping It Simple

The term *productivity* connotes both effectiveness and efficiency. That is, one should take into account both the extent to which the targets have been achieved and the costs incurred in achieving them. Introducing considerations of cost substantially complicates the discussion, well beyond the scope of this short paper. (For a recent comparative analysis of the productivity of different educational interventions, see Belfield & Levin, 2007). Consequently, I will confine my attention to effectiveness.

As noted earlier, there is a clear expectation on the part of the Department of Education that the assessment system should yield evidence to support the determination of the effectiveness of schools, principals, and teachers. The intention is to hold these units accountable for their performance and to provide information that can lead to improved performance. In principle, the notion of holding service units, as well as the individuals in those units, accountable for the use of public funds is unobjectionable. However, implementation of a system that accomplishes this task without causing negative unintended consequences is problematic, to say the least. Again, there is voluminous literature. A particularly thoughtful analysis and discussion of *performance monitoring* from an English perspective is provided by Bird et al. (2005). They note that one purpose of performance monitoring is

> . . . to give the public a better idea of how Government policies change the public services and to improve their effectiveness. Performance monitoring done well is broadly productive for those concerned. Done badly, it can be very costly and not merely ineffective but harmful and indeed destructive. (Bird et al., 2005, p. 1)

That same message of promise and peril (with an emphasis on the peril) is echoed by Rothstein (2008), who conducts a review of the literature on evaluation and performance incentives in other fields, both public and private. He notes the high likelihood of negative unintended consequences stemming from the imposition of high-stakes accountability. Although the current Administration is determined to press ahead, caution and humility are in order.

Investigations of effectiveness usually begin with a focus on one or both of the following questions:

- Where are the students (in a classroom, in a grade, in a school) located on an appropriate scale representing academic achievement?
- How much did the students (in a classroom, in a grade, in a school) learn during the course of the academic year?

It is self-evident that in order to answer these questions credibly, the assessment system must generate information that possesses a high degree of construct validity with respect to both attainment and growth. This is the responsibility of assessment design and development, as well as of administration and data processing. To the extent that the measures suffer from construct underrepresentation, or construct irrelevant variance, there is an increased danger of negative unintended consequences (Baker, 2002).

To move from description to evaluation, a suitable framework must be established. Typically, the framework leads to absolute or normative comparisons. Contrasting schools with respect to the average scores of their students in a particular grade is a normative comparison. Comparing each school's average to a predetermined target is an absolute comparison. The latter is the approach adopted by NCLB. In either case, the criticism is that current status is the result of the cumulative contributions over the life of the child and, consequently, an inappropriate basis for judging the effectiveness of the school in which the child is currently enrolled. This has led to calls to modify the NCLB regulations to permit consideration of measures based on how much students had learned during the year—so-called growth-based indicators.

The cogency of that argument, as well as political pressures, led then Secretary Spellings to approve a compromise, the Growth Model Pilot Program (U.S. Department of Education, 2005), which allows schools to get credit in the current year for students who have not achieved the proficiency standard that year but who are on track to proficiency" within a specified time horizon. Different *growth to a standard* models were proposed and approved. It has been pointed out, however, that schools with lower achieving students still face greater challenges than schools with higher achieving students, since they are not being judged solely on the basis of the amount of progress their students had made during the year.

This brings us to the second question above, which does concern learning *per se*. Answering this question requires longitudinal test records for individuals. Again, for evaluative purposes, absolute or normative comparisons—and even some combination of the two—can be called for. As will be clear in what follows, methods of varying statistical sophistication can be employed. However, in the present NCLB era, and perhaps beyond, test performance in relation to proficiency standards is the coin of the realm. Leaving aside the question of whether a particular proficiency standard can be meaningfully interpreted, an interest in growth will naturally lead policymakers, the public-at-large, and parents to focus on the reported performance of individual students, or groups of students, as they move from grade to grade. In principle, schools could be compared with respect to changes in the percent proficient from one grade to the next.

There are serious concerns with such an approach. As it stands today, for example, one can observe that 60% of students in a cohort were deemed proficient in Grade 3 but only 40% are proficient in Grade 4, and conclude that the Grade 4 teachers were less effective than those in Grade 3. The conclusion may well not be warranted, because the proficiency standard for Grade 4 might be more rigorous than the one for Grade 3. Such incoherence across grades will inevitably cause misinterpretation. Notwithstanding certain technical problems associated with using percent proficient as an indicator, changes in percent proficient will surely remain an attractive basis for evaluation. For this, and other reasons, devising a set of vertically coherent standards should be a high priority.

In this regard, Bejar, Braun, & Tannenbaum (2007) argued that the processes of test development and standard setting should be better integrated and offer some suggestions on how greater vertical coherence could be achieved by adopting a more explicit developmental perspective—one that would be consistent with an emphasis on student learning. They also pointed out that if a set of end-of-high-school standards are in place, then it would be possible to work backwards from those standards to set articulated standards in earlier grades, such that a student achieving proficiency in grade n would be on track (i.e., have a high probability) to achieve proficiency in grade (n+1). In such a case, although the percentages of students achieving proficiency in successive grades would still be a crude indicator of effectiveness, they would be less likely to lead to grossly misleading inferences. Note that such an approach to standard setting would be very much in line with the desiderata of the notice (U.S. Department of Education, 2009).

As attractive as the prospect of coherent standards may be, one should not underestimate the challenge in developing such a system. There are implications for how content standards should be constructed, as well as how assessment development proceeds. Regrettably, there is insufficient experience in carrying through such a program. Note, by the way, that if the end-of-school *readiness standards* are empirically tied to real-world demands, then they serve as a meaningful anchor for the academic standards in the earlier grades. At the moment, each state's standards constitute a *hermetically sealed system*, essentially divorced from the reality checks provided by linkages to the world outside school—which may account, in part, for the substantial variability in rigor of the standards across states. It is also a problem that has long been noted, as the following quote attests:

> The chief fault of the testing movement has consisted of its emphasis on content in highly academic material . . . the fact that a particular pupil shows a

> marked improvement in reading or spelling may give some indication that a teacher is improving her performance . . . but the use to which the pupil puts that knowledge is the only significant point in determining the significance of subject tests in measuring the educational system. (Ridley & Simon [1938], as quoted in Rothstein, 2008, p. 10)

More recently, Haertel and Lorie (2004) made a related point:

> Arguments and procedures supporting a performance standard . . . may differ according to the breadth of the claim the performance standard sets forth . . . . In practice, though, the performance standard always embodies a . . . claim pertaining to capabilities for performance in nontest settings. (p. 63-64)

The difficulty, then, is that too often the claims attached to present day performance standards are essentially rhetorical—they are not supported by the process that generated those standards. The hope is that greater integration of assessment development and standard setting, coupled with vertical coherence and links to real-world tasks, will yield performance standards that are more interpretable and, hence, more useful.

## Growth Modeling

Evaluating effectiveness on the basis of charting a cohort's progress with respect to cross-grade proficiency standards is rather crude, in more ways than one. A seemingly viable alternative is to track each individual's progress along a vertically linked score scale and to construct an indicator based on the average gain across individuals within a unit (e.g., a grade within a school). The indicators could be judged normatively or absolutely. Gain scores are intuitively appealing and attractive because they are more weakly correlated with student demographics than are status measures.

As pointed out above, however, there are problems in relying on a vertical scale for evaluation purposes. Moreover, gain scores can be volatile and induce bias in estimation both because of measurement error and missing data (Ladd & Walsh, 2002). They also do not fully account for differences among educational units on relevant student characteristics.

## Value-Added Modeling

One step beyond the use of conventional gain scores is value-added modeling (VAM). Typically, VAM refers to a class of statistical models that are used to estimate the effectiveness of schools and teachers (Braun, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Employing individual-level longitudinal test data, such models seek to extract a component of aggregate test score trajectories that can be attributed to a particular educational unit. (This is the rationale for the use of the term *value-added*.) The difficulty lies in the fact that students are not randomly allocated to schools or to teachers within schools. Consequently, straightforward estimates can be contaminated by selection bias. VAM attempts to adjust for this selection bias in order to obtain unbiased estimates of the target quantities. This is sometimes referred to as *leveling the playing field*. In this respect, VAM differs from straightforward growth modeling, which does not involve any statistical adjustment.

The problem is that it is very challenging to determine how successful the adjustment process has been in removing selection bias. Thus, many observers harbor serious concerns about endowing causal interpretations of indicators based on data obtained through an observational study rather than a randomized experiment. This explains, at least in part, the ongoing controversy as to whether the results of a value-added analysis should be used for purposes of high-stakes accountability. For a recent summary of the issues surrounding VAM and its applications, see the report by Braun, Chudowsky, & Koenig (2010).

Some of the issues concern the assumptions about the test data that are made by each type of VAM. In addition to the bedrock assumption of construct validity, many approaches assume that scores from different grades have been placed on a single cross-grade scale. As noted earlier, the process of vertical linking is fraught with technical and substantive difficulties (Briggs & Weeks, 2009), with implications for the interpretation of the resulting value-added estimates. Many different models have been used to obtain estimates of value-added, each with its advantages and disadvantages as determined both by theoretical analysis and empirical investigation. Selecting a model that is most suitable for a specific context and purpose is not a straightforward exercise.

It should also be borne in mind that the results of a value-added analysis are almost always expressed in normative terms. That is, the units to be evaluated are ranked in order of their (estimated) relative contributions. For example, many approaches utilize regression models to construct an expected result for each unit. The role of the regression model is to take account of the observed differences among units on variables that are statistically associated with test performance. The difference between the observed result and the expected result (sometimes modified to reduce sampling volatility) is taken as an indicator of the unit's value added. Note that because the regression model is estimated from the data on the same set of units to be evaluated, the value-added estimates are essentially regression residuals. Consequently, it is necessarily the case that about half the units will be assigned positive values and half negative values. Nonetheless, students at units with negative values may still have posted substantial gains.

## Complex Growth Data

The notice (U.S. Department of Education, 2009) called for the design and implementation of an assessment system that will generate more extensive data on student learning for summative assessment than is currently the case. Logically, if federal and state governments are to invest in building these next-generation assessment systems, it is incumbent upon the states to make fuller use of the data they generate. If the costs in time and effort are to be justified, then we need more elaborate notions of growth to fully reflect the learning that is captured by the system. Unfortunately, almost all of the work on test-based accountability thus far has assumed that a student's learning trajectory can be adequately summarized in a temporal sequence of scalar test scores that are amenable to traditional statistical analysis. Evidence is accumulating, however, that the results of an accountability system can vary substantially for two different standardized tests of the same material, even if the tests are similar in format and difficulty (Lockwood et al., 2007). Thus, we should expect even greater variability when more complex sorts of evidence are brought to the table—assuming we know how to build an appropriate table!

Two approaches are available immediately (and surely others can be developed). One is to build a student profile based on the data and treat that profile as a multidimensional vector that varies through time. The education value-added assessment system (EVAAS) model developed by W. Sanders and his associates (Sanders, Saxton, & Horn, 1997) is built to accommodate such profiles, though until now its input has consisted of test scores from different subjects. However, there is no reason in principle why it could not operate on such profiles. Another approach is to build learning progressions (Wilson & Bertenthal, 2006) that capture likely learning pathways and are marked by substantively meaningful developmental milestones. A similar exercise has been undertaken in England, where age-independent learning milestones have been established in various subjects (Wiliam, 2006). Based on theoretical and empirical considerations, a student profile could be linked to a milestone and student learning would be quantified by means of transitions from one milestone to another. Individual transitions could be aggregated to the level of the unit to be evaluated (such as a school), resulting in a transition matrix for the unit. These transition matrices could then serve as the input to a specialized value-added analysis with the results employed to compare units (Braun, Qu, & Trapani, 2008).

## Multiple Indicators

To this point, the analysis has focused on the use of growth measures derived from test scores as a basis for evaluation. Both technical and political considerations point to the likely implementation of a system employing multiple indicators. Strictly speaking, any test-based indicator is fallible—it is affected by both bias and measurement error. In addition, it can capture only some aspects of a unit's educational outcomes. It is doubtful that a single indicator can—or should—bear the burden of supporting high-stakes evaluations. Consequently, relying on a set of indicators can mitigate the risk of making poor inferences. The indicators should be selected to represent different educational goals and each should be subject to a validity analysis. It is also the case that it is more difficult to *game a system* with multiple indicators than one with a single indicator. Thus, such systems should be less likely to generate perverse incentives and be less subject to corruption.

Notwithstanding the appropriate criticisms of using status measures for accountability and the enthusiasm for growth measures, in some quarters there are strong opinions regarding the importance of retaining status indicators in an evaluation system. To cite just one example, among advocates for students with special needs, there is the belief that the pressure exerted by NCLB to have all students reach a common proficiency standard, coupled with the requirement that school data be viewed overall and for each designated subgroup, has led to greater academic rigor in the programs for special needs students and a concomitant increase in achievement. Although they are appreciative of the rationale of incorporating growth measures, they are concerned that sole reliance on growth-based metrics may lead to lowered expectations and a loss of momentum in the gains these students have made (National Center for Learning Disabilities, 2009). Status measures are easy to understand and, it must be said, place a lesser burden on the assessment system than do growth measures. For these and other reasons, multiple test-based measures are likely to feature in the reauthorized ESEA.

This conclusion gives rise to two other issues. The first is what non–test-based indicators could be developed to broaden the evidence base for evaluation. The second is how these different indicators should be combined for purposes of decision making, as well as the allocation of rewards and sanctions.

Neither issue is simple, but they must both be addressed as the state consortia endeavor to meet the federal requirements.

## Technology

In a brief treatment such as this, it is impossible to even touch on all the relevant issues. One such issue is the role of technology. It appears inevitable that most assessment will migrate to technology delivery. The only question is at what pace and at what level. The reason is that there are enormous benefits to be had, for both diagnostic and summative assessment. With computer technology, a broader array of item types can be employed, improving construct validity. Utilizing expert systems of different levels of sophistication, student-constructed responses can be scored automatically and useful diagnostic information provided as well (Williamson, Mislevy, & Bejar, 2006). Moreover, adaptive testing can be employed to improve both precision and utility. Finally, turnaround times can be instantaneous for diagnostic assessment and nearly so for summative assessment.

However, the last decade has shown how difficult it is to realize the potential of technology, although some states and districts have made considerable progress. Clearly, lack of expertise and cost are two major obstacles. Implementation strategies must take into account the varying capabilities at different levels of the education system and plan for the gradual upgrading of relevant capabilities over an extended period. With respect to cost, funds should be invested in building capacities that can be leveraged in different ways. This is explicit in the notice (U.S. Department of Education, 2009), for example, in the expectation that the technology platform will support other related activities. (A technology platform is characterized by a set of mutually compatible components with generic capabilities that can be relatively easily harnessed in various combinations for different purposes.) Thus, a technology platform should be able to support both instruction and assessment. Other dual-use strategies could involve teacher professional development and local-level (or state-level) collaborative communities.

With respect to instruction, technology allows students to utilize various software packages for word processing, spreadsheets, presentations, and the like. Of course, familiarity with these packages is not an end in itself but a means to developing expertise in different content areas. Access to the Internet enables students to link to myriad information sources and to interact with individuals and institutions around the world. Equally exciting, technology can support the creation of virtual worlds in which, for example, students can conduct experiments or construct interacting systems of their own. Ultimately, with technology ubiquitous in the world outside of school, the goals of schooling will come to include capabilities essential for life in the digital age (Zhao, 2009). Thus, technology, like assessment, will serve as both tool and driver for curriculum and instruction—with all the attendant possibilities and pitfalls.

Another issue is the impact of changes in assessment content and delivery on students with disabilities. Considerations of access and equity should influence design decisions from the outset. Principles of universal design and adaptive technology can result in improvements for many students with disabilities, but there may well be some that will be disadvantaged. There are lessons to be learned from admissions testing in higher education where technology platforms have been in use for more than 15 years.

## Discussion

Through its Race to the Top initiative, the Department of Education has set out an ambitious agenda. Two components of this agenda—assessment and accountability—are the subject of this paper. The required characteristics set out in the notice (U.S. Department of Education, 2009) make it clear that the authors have been assiduous in gleaning from the literature many of the qualities that would characterize an ideal assessment system. Moreover, it is expected that the results of these assessments will serve as input to the next instantiation of accountability systems. There is much to admire in this initiative: The requirements that only consortia of states may apply and that their proposals must provide a plan to develop a set of "common summative assessments that are aligned with a common set of K–12 internationally benchmarked, college and career-ready standards." (U.S. Department of Education, 2009, p. 1) could move American education in the right direction while eliminating some of the redundancy and waste inherent in having 50 separate systems.

If this nation were starting with the proverbial blank slate, substantial progress could be made in three to four years. Given current realities, however, it will be more difficult, impossible perhaps, to achieve these goals by the end of the Administration's current term. First, a consortium of states formed under some duress is certain to experience tensions arising from differences in philosophies, values, and preferred approaches to assessment. On a more prosaic level, each state's current procurement policies and contractual obligations differ in scope, timing, and cost. How these differences could be resolved in a relatively short period of time remains to be seen. Moreover, even with the best of intentions among state leaders, there are concerns regarding capacity at state departments of education, as well as the likelihood of both resistance and simple inertia at lower levels.

With these cautions in mind, what is a reasonable but still ambitious and meaningful target? I would argue that a consortium should be expected to chart a new pathway for assessment development that will jumpstart innovation and lead to a first approximation to the ideal. For example, the new system should:

- Model new patterns of collaboration among states and contractors in the design, development, and validation of new assessments;
- Have superior measurement properties with respect to indicators of status and growth;
- Effectively employ new paradigms for comprehensive assessment; and
- Exhibit the potential of technology platforms.

Of course, the context for assessment development comprises not only existing assessment practices and artifacts, but also systems of curriculum, instruction, professional development, and so on. The dynamics among these interrelated systems, further complicated by both political and market forces, will surely influence the nature and pace of innovation. In the face of these (potential) obstacles, there is a natural tendency to be overly prescriptive. This should be avoided, as it may limit creative solutions. Consortia should be given reasonable flexibility in responding to the notice (U.S. Department of Education, 2009), consistent with the general intentions. For example, states within a consortium could

be allowed to adopt the common assessments on different schedules, depending on their contractual obligations and other legitimate constraints.

The interplay between assessment and accountability is especially complex. On the one hand, the goal should be to maximize construct validity (in its various aspects) of the tests within the limitations imposed by the context and practical constraints. On the other hand, a summative assessment system will be embedded in an accountability system that comprises procedures for constructing indicators and decision rules (drawing on those indicators) that often lead to rewards and sanctions. In such a context, it is most useful to consider the validity of the accountability system as a whole, rather than that of its individual components. Indeed, it is its consequential validity that is of greatest import.

One way of representing this aspect of validity was provided by Braun and Kanjee (2006) under the rubric of *systemic validity*. Paraphrasing, they proposed that assessment practices and systems of accountability be considered systemically valid if they generate useful information and constructive responses by educators and education officials that support one or more policy goals (access, quality, equity, efficiency) of the education system, without causing undue deterioration with respect to other goals. More specifically, in the context of public education, the goals would likely include:

- Support good instruction;
- Positively influence teacher quality, recruitment, and retention;
- Promote student access to high quality (and meaningful) educational experiences; and
- Build school capacity through appropriate resource allocation.

The challenge inherent in achieving these goals highlights the importance of making explicit the theory-of-action underlying the design of the accountability system and its components. That is, how will the system accomplish the goals enunciated by the responsible officials? Equally important, what are the possible unintended consequences once the system is put into place—and how does the design preemptively address these potential problems? For example, under NCLB, attention focuses on reading, mathematics, and science. As a result, there is considerable evidence that in many schools the curriculum has narrowed—in two different ways. First, in tested subjects, teachers are teaching to the test, often by drilling students on the item formats found on the test. Thus, when tests rely heavily (or exclusively) on multiple choice items, the content of instruction is greatly limited. Second, less class time is devoted to nontested subjects and some (particularly the arts) are even eliminated from the curriculum (McMurrer, 2007).

If the next generation of assessments and accountability systems induces schools to make content more relevant and instruction more effective, then student development will surely profit thereby. If it fails to do so, then aggregate learning will continue to stagnate and the high dropout rates we now observe will persist (Johanek, 2009). Equally damaging, an evaluation system that is considered to be unfair may hasten the departure of good teachers and discourage those who are considering a teaching career from entering the profession altogether.

With the stakes literally so high, it is essential to plan for an audit of the accountability system so as to document its impact on various facets of education. In this regard, it is important to take account of the

*evidential asymmetry* inherent in such an audit. By that is meant that test scores, as well as certain types of administrative data, are recorded as a matter of course and generally require little additional expenditure to be incorporated in an audit report. However, data that bear on other aspects of education, such as changes in the allocation of resources and school-level decision making, choices in pedagogical strategies, and the like, are more difficult to capture, often requiring specialized studies and longer timeframes. It is precisely this kind of data that can provide evidence of unintended consequences, both positive and negative. Appreciating and accommodating this asymmetry would greatly enhance the value of any audit. For a comprehensive treatment of the standards to which accountability system designers should aspire, see Baker and Linn (2004).

Certainly, the road ahead is a challenging one and the stakes are enormous. However, lessons learned from past successes and failures—and, in particular, the limitations of top-down reform efforts—can offer a sound foundation for constructive progress.

## References

Baker, G. (2002). Distortion and risk in optimal performance contracts. *Journal of Human Resources, 37*(4), 728–751.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47–72). New York, NY: Teachers College Press.

Barton, P. (2009). *National education standards: Getting beneath the surface* (ETS Policy Information Perspective). Princeton, NJ: ETS.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 31–63). Maple Grove, MN: JAM Press.

Belfield, C. R, & Levin, H. M. (Eds.). (2007). *The price we pay: Economic and social consequences of inadequate education*. Washington, DC: The Brookings Institution.

Bird, S. M., Cox, D., Farewell, V., Goldstein, H., Holt, T., & Smith, P. C. (2005). Performance indicators: Good, bad and ugly. *Journal of the Royal Statistical Society A, 168*(1), 1–27.

Braun, H. I. (2000). A post-modern view of the problem of language assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 263–272). Cambridge, England: Cambridge University Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models* (Policy Information Perspective). Princeton, NJ: ETS.

Braun, H. I., Chapman, L., & Vezzu, S. (2010). *Using state NAEP data to examine patterns in eighth grade mathematics achievement for Black students and White students*. Unpublished manuscript, Boston College.

Braun, H. I., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Braun, H. I., & Kanjee, A. (2006). Using assessment to improve education in developing nations. In J. E. Cohen, D. E. Bloom, & M. B. Malin (Eds.), *Educating all children: A global agenda* (pp. 303–353). Cambridge, MA: MIT Press.

Braun, H. I., Qu, Y., & Trapani, C. (2008). *Robustness of value-added analysis of school effectiveness* (ETS Research Rep. No. RR-08-22). Princeton, NJ: ETS.

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black–White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14*(8). Retrieved from http://epaa.asu.edu/ojs/article/viewFile/79/205.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3–14.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*, 67–90.

Donovan, M. S., Bransford, J. D., & Pellegrino, J.W. (Eds.). (1999). *How people learn: Bridging research and practice.* Washington, DC: National Academies Press.

Education Week. (2009, June 11). Diplomas count 2009 [annual report]. *Education Week, 28*(34).

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61–103.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 433–470). Westport, CT: Praeger.

Johanek, M. C. (2009). *School reform that matters.* Retrieved from http://www.gse.upenn.edu/review/feature/johanek#notes.

Koretz, D. (2008). The pending reauthorization of NCLB. In G. Sunderman (Ed.), *Holding NCLB accountable: Achieving accountability, equity, & school reform* (pp. 9–26). Thousand Oaks, CA: Corwin Press.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21*, 1–17.

Lee, J. (2007). *The testing gap*. Charlotte, NC: Information Age Publishing.

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007-494). , Washington, DC: National Center for Education Statistics.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L., Stecher, B., Le, V-N., & Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing*. Charlotte, NC: Information Age Publishing.

Martineau, J. A. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35–62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McMurrer, J. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Retrieved from http://www.cep-dc.org/index.cfm?fuseaction=document.showDocumentByID&nodeID=1&DocumentID=212.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

National Center for Learning Disabilities. (2009). *Growth models for accountability: Considerations and recommendations for including students with disabilities* (Policy Briefing). Washington DC: Author.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Rothstein, R. (2008). *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education*. Nashville, TN: Vanderbilt University, National Center on Performance Incentives.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluational measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

U.S. Department of Education. (2005). *Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond* [Press release]. Retrieved from http://www.ed.gov/news/pressreleases/2005/11/11182005.html.

U.S. Department of Education. (2009). *Executive summary, Race to the top assessment program: Notice of public meetings and request for input*. Retrieved from http://www.ed.gov/programs/racetothetop-assessment/executive-summary.pdf.

Wiley, D. E. (2002). Validity of constructs versus construct validity of scores. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 207–227). Mahwah, NJ: Lawrence Erlbaum Associates.

Wiliam, D. (2006). *Once you know what they have learned, what do you do next? Designing curriculum and assessment for growth*. Unpublished manuscript, University of London, London, England.

Williamson, D. M., Mislevy, R.M., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. B., & Bertenthal, M. W. (Eds.). (2006). *Systems for state science assessment*. Washington, DC: National Academies Press.

Zhao, Y. (2009). *Catching up or leading the way*. Alexandria, VA: ASCD.