

# **Finding Additional Value in New Accountability Systems**

**December 2004**

**Center for Educational Decision Support Systems  
North Central Regional Educational Laboratory**

**Arie van der Ploeg**  
*Learning Point Associates*

**Yeow Meng Thum**  
*University of California at Los Angeles*



1120 East Diehl Road, Suite 200  
Naperville, IL 60563-1486  
800-356-2735 • 630-649-6500  
[www.learningpt.org](http://www.learningpt.org)

Copyright © 2004 Learning Point Associates, sponsored under government contract number ED-01-CO-0011. All rights reserved.

This work was originally produced in whole or in part by the North Central Regional Educational Laboratory with funds from the Institute of Education Sciences (IES), U.S. Department of Education, under contract number ED-01-CO-0011. The content does not necessarily reflect the position or policy of IES or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by the federal government.

Learning Point Associates was founded as the North Central Regional Educational Laboratory (NCREL) in 1984. NCREL continues its research and development work as a wholly owned subsidiary of Learning Point Associates.

## Introduction

The No Child Left Behind (NCLB) Act states a clear purpose, to “ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic standards and state academic assessments” (Sec. 1001). It sets a clear target: by the end of the 2013–14 school year, all public school students “will meet or exceed the State’s proficient level of academic achievement” (Sec. 1111b [2] [F]). This purpose and this target are laudable.

The law requires each state to develop a monitoring and accountability system to measure that targets are being reached. Critical to this system are the adequate yearly progress (AYP) criteria. AYP defines, uniquely within each state, a sequence of performance benchmarks for the state, its districts, and its schools. The benchmarks rise with time. Each year, a school—not a student—meets or does not meet the benchmark. Schools are successful or not; sanctions apply to schools, not students.

NCLB’s purpose and target speak of students; AYP speaks to schools. This shift has numerous implications. This paper addresses some. Its core argument rests on the observation that while NCLB’s AYP-based accountability system addresses school performance, for schools to make the progress being asked of them, they will need guidance to improve teaching and learning. The blunt fact is that knowing that a school does not meet AYP is not informative about how to improve teaching practice.

Put another way, for school personnel the central issue posed by NCLB accountability is finding viable solutions to questions like: *Given where we are now, are we improving at a rate that will keep us on track to reach the target in the time remaining? If we are improving too slowly, what must we do differently?* Useful answers to such questions will speak to what teachers and students do together over the 180-or-so days of the school year, in particular to what they should do differently. Is it possible to augment state accountability systems so that they speak more to such matters?

The authors conclude that this is possible. The answer is in two parts: an introduction (this paper) and the response (a subsequent paper in early 2005). This paper addresses some basic issues and introduces several others that structure thought about how new accountability systems can generate additional utility. The subsequent paper will detail a schematic for the design and implementation of appropriate solutions given the knowledge, technology, and capacity already available.

States, districts, and schools are today implementing new and better mechanisms for collecting, storing, and manipulating data. These systems are electronic and often Web-based. Over the next few years, more data and more accurate data will become readily available.<sup>1</sup> Statewide student

---

<sup>1</sup> A variety of public, private and joint ventures are making this happen. One example is the U.S. Department of Education’s collaboration with the Broad Foundation which funds a partnership between Standard and Poor’s School Evaluation Services and Just for the Kids that delivers Web-based access to school performance data nationally (see the Web site <http://www.schoolresults.org>). Another is PBDMI, the Performance-Based Data Management Initiative funded by the chief information officer of the U.S. Department of Education and operating through the Council of Chief State School Officers and ESP Solutions Group. For ongoing status reports on this work, go to <http://evalsoft07.evalsoft.com/pbdmi/>). A third is the Data

identification numbers will permit data records to follow students from school to school. Annual testing of all students in Grades 3–8 in reading and mathematics, and once in high school (as well as in science at three grade levels) will be the norm in all 50 states. Districts and schools will connect (in fact, many already are) other data to this core, including student demographic and program characteristics, grades, course enrollments, and attendance. Local capacity for data-driven decision making about teaching and learning will grow rapidly. So far, however, few states have given careful thought to how to encourage this expanded local capacity as they extend their own data systems.<sup>2</sup>

No matter how sophisticated, an accountability system will be no more valid than the measures on which it is based (Hill & DePascale, 2003) and no more useful than the numbers it reports. NCLB and state accountability systems both rely heavily on statewide standardized tests.<sup>3</sup> The technology of standardized testing is highly sophisticated. Still, three points require recognition: (1) measurement is never perfect—all data contain error; (2) no standardized test completely measures what a student knows; and (3) whether the particular test in use is the “right” test is to some degree always an open question.

These issues are usually left to technical experts:

- Data system engineers and database specialists take responsibility for building secure, safe, efficient systems that store and manipulate important data. Their conversations focus on software and hardware systems, permanence, storage, confidentiality, reliability, throughput, report design and distribution. They talk a language with words like SQL, CICS, J2EE, XML, SIF, and APIs. These conversations are impenetrable to the uninitiated. Prestige accrues in these professions as problems in ever larger, ever more complex systems are solved.
- Psychometricians and statisticians talk about reliability, validity, and inferential risk. They make use of highly sophisticated quantitative tools and obscure acronyms: item response theory (IRT), differential item functioning (DIF), Monte Carlo simulation, hierarchical linear models (HLM), ANOVA, MANOVA, LISREL, Type II errors, the list goes on. What are they talking about? Good measurement and safe decision making. In practice, their work favors large-scale applications because that is where their tools offer greatest traction—not to mention greater funding potential.
- Administrators and practitioners lead schools and teach students. We all know schools, we think. Still, practitioners’ talk also is often impenetrable, jargon laden. Using data to drive decisions is a major current theme. However, the data they most want to use only rarely draw the attention of the other two conversations. Their own attempts to make good use of gleanings from those conversations are hampered by lack of technical skill.<sup>4</sup>

---

Partnership, a collaboration of Achieve, the CELT Corporation, the Council of Chief State School Officers, and Standard and Poor’s, funded by the Broad and Gates Foundations.

<sup>2</sup> See the previous paper in this series by Palaich, Good, and van der Ploeg (2004).

<sup>3</sup> Nebraska’s STARS system is a partial exception (see Roschewski, 2004). However, that system should be even more concerned about the issues raised here; it is still working out its technologies.

<sup>4</sup> As former U.S. Secretary of Education Chester Finn once put it, “Most of the data we need, we cannot get. Much of what we get, we cannot trust. Of that which we can trust, far too much is obsolete, unintelligible to laymen, or unsuited to crucial analyses and comparisons” (Finn, 1991, p. 263). Supovitz and Stein (2004) were “shocked to observe the limited technological capacity of . . . innovative data-using schools.”

None of these specialists working alone can comprehensively address the design and construction of a broadly useful accountability system. This paper suggests that if these specialists will learn to speak effectively with each other, accept certain core design criteria, and agree to design and build together, then enhanced data capacities and technologies will soon extend state and district accountability systems to speak more meaningfully to classroom teachers and school leaders. We begin with several core criteria. In doing so, we draw upon Thum (2003) who recently summarized the methodological literature on measurement in education and identified critical choices to be made when constructing systems to measure academic growth.

### **What Matters—What a Student Knows or Whether a Student Learns?**

What is it that accountability systems monitor? Such systems have a choice: they can tell about status—where students are, what students know—or about change and growth. Both are important; in fact, they are complements.<sup>5</sup> State accountability systems have often favored growth: good schools are schools that improve. NCLB’s accountability system requires annual reports of status: good schools are schools that score above benchmarks.<sup>6</sup> States and districts have worked diligently to meld their own accountability requirements with those of NCLB (Porter, Chester & Schlesinger, 2004).

Understanding the change in a student’s (or group of students) knowledge and ability over time is critical to determine strategies for improvement. The details of how much students grow over a year compared to what was taught during the year help us to see what works and what needs to change. Learning is cumulative and the rate of learning is variable. Knowing only a score at the end of the year, we cannot know what was learned when or how well. A “value-added” approach (Meyer, 1996) to assessing student learning measures the change in students’ knowledge and skills over time and the variations in growth among students.<sup>7</sup>

Measuring change is considerably more difficult than measuring status. At the very least, twice as many data points are needed: a starting value just prior to instruction and a data point immediately after instruction. To measure well, we will need even more data. Data collected *during* instruction (between the starting and end points) will increase the resolution of what we can see. At least one data point taken well *before* the start of instruction will help locate the observed trend lines (that connect the start and end points).

These facts foreground several design issues:

---

<sup>5</sup>The National Center for the Improvement of Educational Assessment was among the first to clarify the complementarities and tensions of these perspectives (e.g., Gong, 2002). There are subtle issues and choices here, some with major consequences. For instance, growth may be viewed as improvement of scores between successive cohorts (e.g., 2004 third graders compared to 2004 fourth graders) or longitudinally (2004 third graders compared to 2005 fourth graders). Results will differ, and so may conclusions about what actions to take. Seltzer et al. (2003) offer an informative discussion.

<sup>6</sup> AYP raises the bar continuously until the benchmark becomes all students proficient at the end of the 2013–14 school year. NCLB’s accountability system requires growth, but does not measure it. Schools can be but are generally not rewarded for improvement; they are sanctioned for not meeting benchmarks.

<sup>7</sup> Raudenbush (2004) demonstrates that value-added approaches provide stronger answers than approaches that focus on mean proficiency. Still, the value-added approach is not bias-free.

- New data systems will need to store far more data than typically expected. Annual test data and identification attributes will not be sufficient. Multiple years of data will be necessary.<sup>8</sup>
- Individual student records, linked longitudinally, will need to be stored. Storing only institutional aggregates, even for all the NCLB disaggregations and other student groupings, will not support the analyses needed.
- Central and local data stores will be needed, and they will need to interconnect. To understand how changing instruction affects changing performance, data on participation, on performance during the instructional process, on the organization and pacing of instruction will need to be factored in. These kinds of data are only available locally.<sup>9</sup>
- The two basic issues—status and growth—will generate multiple corollary questions, in part because the issues are complex, and in part because the data store is rich. Hence, the systems must be designed so as not to curtail questioning early. At the least, they must permit efficient building on of additional functionalities.

### **Accept Error**

A test score is a useful estimate. A test score represents explicitly defined evidence of performance obtained in an operationally consistent manner. However, despite operational consistency, variation creeps into test results from various sources, including measurement imprecision and sampling variations. In addition, no test score can capture all a student knows and can do. In fact, we often have difficulty saying specifically what it is that a test represents. Just what is “reading comprehension”? Being able to read, yes, but in the details this is difficult to define. “Height,” “speed,” or “density” all have concrete, specifiable definitions. The National Institute of Standards and Technology (see [www.nist.gov](http://www.nist.gov)) maintains the standards and benchmarks against which the measurements of science and technology are calibrated. There is no similar bureau for educational concepts and measures.

The technology whereby we generate tests and score them is immensely sophisticated. The nature of the error structure of soundly implemented testing programs is well understood. We know that random factors other than what students know and are able to do affect test scores. These factors and others contribute to the difference between a student’s “true” score—the hypothetical accurate portrayal of what a student knows and is able to do—and the student’s obtained score. This imprecision is captured by a test’s standard error of measurement (SEM). Most consumers of test data are unaware of how large (or small) a test score’s SEM is; most reports of test score data do not even mention it. Yet, without it, the precision of the reported result is unknown and any decision made will be made with unknown risk.

To repeat, a test score is a useful estimate. As with all estimates, it carries error; we need to be alert to the precision of our measures. We accept that our own weight will vary somewhat from day to day and scale to scale. We learn—typically from experience—how much variation is tolerable before we purchase a new scale. When considering student test results, we also need to

---

<sup>8</sup> A recent systematic comparison in a large urban school system of alternative value-added procedures found that at least five years of data were required to produce “respectable” indicators of school improvement (Raudenbush, 2004, p. 33).

<sup>9</sup> One argument for local systems to extend state and NCLB accountability systems stresses that local systems are created and designed to guide reform over time while state systems simply point to need (see Crane, Rabinowitz, & Zimmerman, 2004).

learn how much variation in the results we can permit before deciding the result is anomalous or the measurement flawed.

However, school averages from large-scale assessments “bounce” markedly from year to year. Volatility is high, even for large groups. Differences occur in the performance of cohorts of students tested in successive years, even when the students are drawn from the same families and the same neighborhoods each year (Kane & Staiger, 2002; Linn & Haug, 2002). Because the average elementary school contains fewer than 70 students per grade, plausible explanations like several unusually successful students, a few students with undiagnosed disabilities, or a particularly effective teacher one year are offered but not demonstrated. In point of fact, no persuasive explanation has been offered for this volatility. Absent an explanation, there is serious risk in judging performance and even more so in recommendations for change in instruction.

Accountability systems report a variety of calculations, projections, comparisons, rankings, and categorizations of performance and productivity. When not accompanied by explicit accounts of the precision of the various measurements taken, procedures used, and sampling variances present, these reports carry a dangerously false sense of security. High-stakes decisions should always be accompanied by carefully constructed statements that fully and clearly represent the extent to which the information is usable and actionable.

There exists an understandable desire to keep accountability systems lean: the fewer the data elements reported, it is argued, the clearer the results, the less the confusion. However, for accountability in education, this often turns out not to be true. Redundancy of information from multiple measures reduces the risks of inadvertent and gross misjudgments.

Measurement error and measure imprecision have clear design consequences:

- To offset the imprecision of single measures, use of multiple measures which overlap each other conceptually enriches the construct assessed and adds strength to results. That is to say, in a world of imperfect measures, redundancy is a good thing. (And, of course, the data store is again enlarged.)
- Simultaneous analysis of multiple measures requires more complex structuring of the data and the application of more sophisticated statistical processing. The information-carrying capacity of rational redundancy hand-in-hand with a well-structured analytic approach promises higher resolution of results and clearer specification of consequences.
- The result to be interpreted, the one that is potentially actionable, may not appear in the first-order accounting of current scores but rather in the second-order trend or pattern across multiple elements simultaneously. That is to say, it may be better to look at the trends of scores than at the value of today’s score. More so, it may be even better to look at the trend of *estimated* scores than at the trend of *actual* scores. If there is “bounce” in the data, then there is value in identifying the pattern hidden behind the bounce.

## **Metrics Matter**

To measure change in student knowledge and ability, we need tests that tap the same construct and measure on the same scale. To be able to see how much learning changes in time, we need to

compare learning across multiple grades. The ruler we use to understand test results from Grade 3 should be the same ruler we use to understand the test results from Grade 4. If we use a different ruler each time we test, determining the amount of progress students make is made more difficult.<sup>10</sup> There are at least three critical implications here for the tests accountability systems use.

The first is consistency of meaning. If we are testing reading, then the meaning of a reading result for a Grade 3 student must be interpretable in terms of a Grade 5 reading score. This implies that in constructing the testing program, the developers and their curriculum allies must share a consistent view of what reading is and how expertise in reading varies over the grades. This view must be made explicit, written down, and broadly disseminated. If it is not written and widely disseminated, the content of teaching cannot be related to the content of the test.<sup>11</sup>

The second is consistency of measurement. At an instrumental level, it should be possible to make statements like: this eighth grader knows three times as much, or 43 ‘units’ more, than that fourth grader. We can do this with height, weight, or temperature, but not for most school assessments—at least not in terms of the numbers in which they are commonly reported. The metric of a test, the units and numbers we use in our analyses, is the critical foundation for our decision making (Seltzer, Frank, & Bryk, 1994). Academic performance measures, like rulers, need to possess equal intervals: an inch should be an inch, regardless of location, time, and temperature, whatever.<sup>12</sup>

The third is consistency of range of measurement. The tests should measure using units (or intervals) that are uniform and that span the full range of the construct being measured. Absence of the construct translates to a “0” on the scale. Small changes for those just gaining skill in the construct must be just as measurable as small changes among those highly skilled.

Similar consistency does not exist in many of the scales and measures we use in schools. Thum (2003) suggests, “It is the responsibility of the test producers to continuously provide the necessary evidence for their scales, making explicit any shifts in procedures or assumptions ..., so as to support the appropriate use of their scales.” This injunction targets state, district, and school staff making and using assessments just as much as it targets commercial test developers. Unfortunately, few data systems that store and report assessment and other data observe this injunction. Consequently, when users interpret the reports that accountability systems provide, they may make comparisons that are unwarranted given the limitations in the measures and scales used.

The NCLB accountability system is a case in point. It requires reporting only the percentage of students who score at a proficient level or above. This dichotomizes every student’s result.

---

<sup>10</sup> It is, of course, possible to equate measures. Mark Reckase (2004) provides an accessible and succinct description of how different but related constructs may be measured independently but scaled commonly. However, this is not a simple task and is fraught with challenging (and often underestimated) substantive and instrumental issues.

<sup>11</sup> The current standards-based instruction movement goes a long way in this direction, but the work is not finished. Frequently, classroom teachers remain uncertain of the practical meaning of the standards for their daily work. Nor are they always convinced that state assessments align well to the standards or to the instruction they believe they are expected to deliver.

<sup>12</sup> At minimum, if intervals are not equal there must exist mathematical transformations that create equal intervals, similar conceptually to how recipes are systematically adjusted for differences in altitude.

Whatever the observed score, only the fact that the score is above or below the proficiency cut-off survives.<sup>13</sup> This clearly throws away much valuable information about a student's learning. Many changes in student performance will not cross a proficiency cut-off value. That does not mean a student's improvement was small. A fifth grader who moves from not reading to reading has made major progress, even if she remains far behind her age peers and is ranked "not proficient" on the fifth grade state test. However, under NCLB reporting rules, her meaningful growth is invisible and irrelevant.

This is not to say that NCLB's focus on counting proficient students is not useful. Clearly, the contrary is true. This metric brings to the fore critical performance issues and group differences. Its value for its own purposes in no way detracts from the value schools, districts, and states can and should draw from the continuous, equal-interval scales their testing programs generate. There is no requirement within NCLB to restrict instructional decision making to the limited information provided by the proficiency reports.<sup>14</sup> In fact, it is not difficult to envision a system that performs accountability analyses using the high-information scale scores and then reports results in terms of proficiency categories.

Some of the design issues that this discussion of data metrics leads to are:

- Educational measurements generate multiple metrics. The data systems to be built must store the required reporting metrics; they also should store the more detailed metrics that assessment systems capture which are not required reporting elements. The system should report in the required performance categories by default. The system should simultaneously recommend and enable reviews of the continuous score metrics.
- Determining which metric is appropriate to a particular question may be a complex substantive or technical issue. For instance, counts of successful achievement against a statewide proficiency cut-off will not usefully display performance differentials in a low-performing district. The data system should therefore permit some variable reconstruction and computation capacity in order to enhance resolution.
- The system should store current information on the quality of each assessment instrument with respect to domain definition, scoring algorithms, score range, validity, precision, reliability, and vertical and horizontal equating.
- The system should recognize the attributes of the variables and metrics stored so that the likelihood of inappropriate data manipulations is reduced.<sup>14</sup> If the system incorporates interactive analytic capacity, it should possess a rules base that links variables and analytic utilities.
- Many of the measures used in education are either approximations (e.g., free lunch counts as a proxy for poverty) or very sophisticated (e.g., normal curve equivalents and growth

---

<sup>13</sup> In addition to the point made in this paragraph, there is a second point, one with several flavors: Do we in fact know what the cut-off point represents? In some sense, it is expected to mean *proficient*. "Proficient in what?" is a fair question since we know the test cannot possibly cover the full domain we specify. We need a description of what a proficient third grader can do. We also assume that proficiency in fourth grade is in some way comparable to proficiency in fifth grade. If a student is not proficient in fourth grade but is proficient a year later in fifth grade, can we say factually and truthfully that the student made progress? Beyond the issue of whether the cut-points in each grade were appropriately determined and operationalized lies the *meaning* question: Does the underlying construct remain the same despite the different operational descriptions of proficiency?

<sup>14</sup> A basic example is prohibiting arithmetic operations on percentile ranks.

indices). In both cases, simplistic analysis is unlikely to return much value. The more sophisticated analytic tools now widely used in academic research need to be brought into the world of schools and districts, to offset the coarseness in the first case and to maximize the information return of the second case. This requires that data system designers understand the record identification and data nesting requirements that support these tools.

- Clearly, an accountability system that supports interaction and local analysis, that brings new analytic tools to unprepared staff, must face the necessity of training its users.

## **From School to Student Performance**

The story that data tell about a school's performance can be very different from the story they tell about student performance, sometimes irreconcilably so, as data are combined and grouped in various ways. Aggregation, or disaggregation, impacts our conclusions by (re)defining the unit being measured and what change is tracked. It is easy to become a bit confused, particularly if our questions are not specific enough or if the statistics we focus upon no longer match the question. For example, should we compare last year's Grade 4 results to this year's Grade 5 results to determine if learning improved? Or should we compare Grade 4 last year to Grade 4 this year? The two questions appear to target the same issue; however, they make use of different data elements, from different students, working with different teachers. Being clear in the purpose of the questions we ask is critical if we are to understand the results obtained.

All approaches to using data to guide decisions presume a model, a well- (or ill-) defined set of relationships that takes us from the data we determine relevant to the decisions we make (Thum 2003). Today's state and NCLB accountability systems leave much of their models implicit, so the user remains unsure how to connect the dots that lead from data to decision. If accountability systems exist to support improvement in schools, then the data they publish must provide guidance for teaching and learning in terms a teacher can work with.

Teachers have expectations about how their classroom activities influence student learning. They carry within them a "model" about how their instruction works. Typically, these "models" are qualitative rather than quantitative, diffuse rather than precise, but nevertheless strongly held. Currently, few accountability systems make any attempts to connect to teachers' personal models and so they are unlikely to influence local decisions. The data disaggregations and reporting that accountability systems provide should reinforce a vision for teaching and learning.<sup>15</sup> An accountability system must make explicit the logic model driving it and present strong arguments aligning its internal statistical processing with its data processing and procedures. Only if the accountability and data logic models are visible, aligned, and understood can schools and teachers determine if these systems and the data flowing from them support or refute their own understandings of how things work and why things change.

---

<sup>15</sup> This point is often not well understood by those who design data systems. Data analysis is purposive, and those purposes should shape how an analysis proceeds. Among other implications, this implies that not all data are equal and that not all ways of looking at data are reasonable or productive. Data systems architects are trained as specialists in data storage and recall and display. They are not trained to think about the purposes for data; they are given "business rules" to implement. Unfortunately, the business rules that drive teaching and learning within schools lack clarity; how we decide what is a good school remains subject to contention.

At the heart of teachers' thinking about teaching and learning is the developing student. For an accountability system to carry weight with teachers, it must speak about individual student growth (in addition to school performance). Teachers do not restrict their view of student growth to the change between two moments in time; they see the student as growing and changing throughout her or his education. Nor do they think only about the student's progress on one subject at a time. Teachers know the child is more complex than that. Data systems and data analysis should carry similar breadth, depth, and richness.

Defensible statistical methods that support longitudinal, multivariate, simultaneous, nested analyses of group and individual performance are being explored.<sup>16</sup> These methods attempt to use all available data at once, thus more realistically summarizing the information available from imperfect data. Because they focus on individual student growth trajectories, they more adequately control for confounding factors. One advantage is that each student serves as his (or her) own control: history is explicitly accounted for, one student at a time. The overall curvature of a student's growth trajectory becomes a useful index of learning change. Using these methods, certain individual student characteristics, such as ethnicity or free-lunch eligibility, no longer bias estimates of the gains students make, although in any one setting it is an empirical matter whether some of these characteristics predict gains. Rather, the instructional experiences fostered in classrooms and schools become the source of relevant explanatory variables. Given evidence that instructional experiences have differential effects, teachers change their practice.<sup>17</sup>

These tools make demanding assumptions and enforce strong requirements on both data and users. They attempt refined answers to very specific questions. The functional and logical relationships among data elements tightly constrain logic and inference, method, and conclusions. This precision is responsible for their value. The assumptions and requirements form a statistical model. To use these tools, we must accept the models and their specifications about how variables are allowed to interact: the statistical models enforce their own rigor about data quality and inference. That rigor must be matched in our mental processing. If that methodological rigor is not maintained, the tools fail.

The requirements of the statistical models enforce a new rigor in our thinking—and teachers' thinking—about educational phenomena. Constructing an account of educational, instructional, and accountability relationships before applying one of these tools is critical. Our understanding of what we need to know and the evidence that we need to know it is sharpened. Our thinking is clarified by the need to face and make such choices. These tools require principled knowledge. They force us to think hard and clearly about our questions and the evidence chain to support a finding. And that is a good and necessary result, although hardly an easy one.

Several design issues are apparent from the foregoing discussion:

- The data systems that undergird state accountability systems should maintain individual student data over at least the full K-12 career, and preferably longer..

---

<sup>16</sup> See for instance Betebenner (2004), Bryk et al. (1998), McCaffrey et al. (2003), McCall, Kingsbury & Olson, 2004), Millman (1997), Raudenbush (2004), Thum (2002), Willms (2000) and Zvoch & Stevens (2003). Despite vociferous debate, a consensus is being reached that these complex approaches provide more powerful analysis and suffer fewer biases (Olson 2004).

<sup>17</sup> Middle school teachers from Pennsylvania's DuBois Area School District, for instance, realized they were pacing instruction too slowly, and offering students too little challenge (Olson 2004, p. 15).

- The data systems should maintain an open interface to district- and school-level data systems, so that data exchange and incorporation are enabled.
- Accountability system analytics should include utilities to display time-based performance trajectories for individuals, groups, grades, and schools.
- Documentation of data systems, data elements, accountability procedures, reporting functionalities, and their underlying logic models must be greatly expanded.
- Training in use of these enhanced functionalities will require significant new budgeting and staff allocation.

### **Leverage Targets**

A target is a valuable benchmark for an accountability system or a school improvement plan. Knowing where we are and knowing the target permits us to subtract the first from the second, thus telling us how much we have to get done. If we also have a date for the target, then we can divide the time to determine how much to get done each year (or week). That simple step has marvelous effect on marshalling energy.

Knowing what needs to get done within each time period enables a monitoring system. We can know whether we are on track to make the target, or not. Statistical methodologies exist that permit continuous evaluation of rates of progress toward future targets relative to some performance baseline (Thum, 2002, 2003). They tell us the likelihood that a particular school's pattern of scores observed over time will meet a target by some later point in time, and how that likelihood changes in time.

When a district sets out on a new course of action, not all schools start at the same time or from the same starting position. Although the same target may be set for all schools, their rates of progress will vary and continue to vary as time passes. Evaluation of the effectiveness of a new course of action needs to know when the new course began and to whom it applied. The measure of the productivity for the new course of action should not include time that elapsed before implementation; that would underestimate productivity. Modeling the locations in time when rates of progress change identifies empirically when schools (or students) begin to accelerate. Accumulating local knowledge of when accelerations in learning occur, for whom, and under what conditions promises stronger evidence-based decisions about programmatic and instructional choices.

### **Insist on Transparency**

Accountability must be transparent, even if the technologies of assessment, data architectures, and analysis are complex. Not everyone needs to open the black box of an accountability system to use its results, just as not every automobile driver needs to understand how a differential works in order to drive safely. It is, however, necessary to leave the key and the manual accessible, right on top of the black box. Someone must be professionally charged with design and safety and be provided access. Access should be open. Certainly, intellectual property needs to be protected. Proprietary systems have a role. Still, although proprietary, these systems serve social ends: their output shapes the future development of individuals. Sound peer review,

professional evaluation, and systematic audits of core processes and operating principles will assure accuracy, comparability, and efficiency while limiting harm from errors in deployment. As these systems receive more realistic road tests, openness assures that inadequacies of method or myths regarding practice will be identified and surmounted.

## Conclusions

NCLB gives increased salience to accountability data and the decisions made from them. It does not specify strong conceptual models to guide teaching and learning or to guide data analysis and interpretation. It leaves the work of providing the necessary models to states, districts, and schools.

It is time to think seriously about taking advantage of NCLB and the developments it has wrought and will continue to bring. Meeting the letter of the law's reporting requirements will not bring all students to 100 percent proficient by the end of the 2013–14 school year. Newer statistical approaches maximize the information that resides in large-scale longitudinal assessment and other data sets. Psychometricians build ever better assessments. The technologies for storing, manipulating, and distributing data have grown wonderfully over the past two decades. Statisticians have solved numerous problems so that well-modeled, deeply detailed, longitudinal, multivariate, suitably nested estimation is now feasible. It is time to bring these capacities and their specialists together to construct jointly new, stronger, more useful applications that support improvement in teaching and learning.

- **NCLB and AYP require reporting of proficiency.** This is useful, but limiting, and insufficient to generate guidance to accelerate learning. The newer data systems that states and districts are creating to meet NCLB reporting requirements should store and support a richer range of data at deeper levels of detail. Building locally valuable displays from these more detailed and more informative data promises great returns for effective instructional decisions.
- **These more detailed data and their analytic results should be represented at the individual student level.** This would enable the system to track progress, and to pinpoint the key points where learning changes in important ways. Teachers can only take advantage of the deep knowledge they possess about their students if they can connect that knowledge to the analytic results and interpret them jointly. That conjunction—individualized data in context with broad personal knowledge—is fertile ground for enhanced understanding of instructional effectiveness, one child and one classroom at a time.
- **Specialists need opportunity to talk meaningfully with one another.** The various specialists involved in designing, constructing, and implementing an accountability system must understand each other's capacities and constraints well enough to assure mutually reinforcing development. The methodological and procedural models that each uses must align with those of the others. The resulting accountability system must present a conceptual and visible model that specifies how instruction is driven and learning accomplished.

- **Every analysis of data involves a model.** NCLB leaves the construction of the necessary models to schools, districts, and states. Principled thought in schools and classrooms about how and why the work there makes a difference will lead to refinement of local goals and practices. Carefully crafted data systems can provide the feedback necessary to identify where improvement is needed and which revised practices work. The stronger the local model, the more the decisions it drives will be evidence-based, the more productive the work of teaching and learning becomes, one classroom at a time.
- **NCLB's goals are set high; however, they need not discourage effort.** Tools to determine targets that effectively challenge students, teachers, and schools have been designed and can be grafted into new data systems that connect state and local data. Better local targeting, better monitoring, customized training delivery, all fitted to local conditions—these give promise of controlled, sustained improvement of student and school performance.

## References

- Betebenner, D. (2004). *An analysis of school district data using value-added methodology*. (CSE Technical Report 622). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles.
- Bryk, A., Thum, Y., Easton, J., & Luppescu, S. (1998). *Academic productivity of Chicago elementary schools*. Chicago: Consortium on Chicago School Research.
- Crane, E., Rabinowitz, S., & Zimmerman, J. (2004). *Locally tailored accountability: Building on your state system in the era of NCLB*. [Knowledge Brief] San Francisco: WestEd. Retrieved December 9, 2004, from [http://www.wested.org/online\\_pubs/KN-04-01.pdf](http://www.wested.org/online_pubs/KN-04-01.pdf)
- Finn, C. (1991). *We must take charge: Our schools and our future*. New York: Free Press.
- Gong, B. (2002). *Designing school accountability systems: Towards a framework and process*. Washington, DC: Council of Chief State School Officers.
- Hill, R., & DePascale, R. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice*, 22(3), pp. 12-20.
- Kane, T., & Staiger, D. (2002). Volatility in school test scores: Implications for test-based accountability systems. In *Brookings Papers on Education Policy 2002* (pp. 235-269). Washington, DC: Brookings Institution.
- Linn, R., & Haug, C. (2002). Stability of school building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), pp. 29-36.
- McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.
- McCall, M., Kingsbury, G., & Olson, A. (2004). *Individual growth and school success*. Lake Oswego, OR: Northwest Evaluation Association.
- Meyer, R. (1996). Value-added indicators of school performance. In E. Hanushek & D. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197-223). Washington, DC: National Academy Press.
- Millman, J. (ed.) (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.
- Olson, L. (2004). Researchers debate merits of "value added" measures. *Education Week*, 24(12, November 17), p. 14-15.
- Olson, L. (2004). "Value added" models gain in popularity. *Education Week*, 24(12, November 17), p. 1, 14-15.
- Palaich, R., Good, D., & van der Ploeg, A. (2004). State education data systems that increase learning and improve accountability. *Policy Issues*, 16(June), pp. 1-11.
- Porter, A, Chester, M., & Schlesinger, M. (2004). Framework for an effective assessment and accountability program: The Philadelphia example. *Teachers College Record*. 106(6), pp.1358-1400.
- Raudenbush, S. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* (William H. Angoff Memorial Lecture). Princeton, NJ: Educational Testing Service.

- Reckase, M. (2004). The challenge of documenting student learning. Presentation at the Council of Chief State School Officers' National Conference on Large-Scale Assessment, Boston, MA.
- Roschewski, P. (2004). History and background of Nebraska's school-based teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice*, 23(2), pp. 9-11.
- Seltzer, M., Choi, K. & Thum, Y. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insights into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis*, 25(3), pp. 263-286.
- Seltzer, M., Frank, K., & Bryk, A. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, 16(1), pp. 41-49.
- Supovitz, J., & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement*. Philadelphia, PA: Consortium for Policy Research in Education (CPRE), University of Pennsylvania.
- Thum, Y. (2002). *Measuring Student and School Progress With the California API* (CSE Technical Report 578). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles.
- Thum, Y. (2002). *Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis*. Santa Monica, CA: Milken Family Foundation.
- Thum, Y. (2003). *No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress* (CSE Technical Report 590). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California at Los Angeles.
- Willms, J. (2000). Monitoring school performance for standards-based reform. *Evaluation and Research in Education*, 14 (3 & 4), pp. 237-253.
- Zvoch, K. & Stevens, J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives*, 11(20). Retrieved December 9, 2004, from <http://epaa.asu.edu/epaa/v11n20/>.