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

Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability

Damian W. Betebenner and Robert L. Linn

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.

Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability

Damian W. Betebenner

National Center for the Improvement of Educational Assessment, Dover, New Hampshire

Robert L. Linn

University of Colorado at Boulder

This paper was presented by Damian W. Betebenner and Robert L. Linn 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.

What Is Growth and Why Measure It?

The widespread availability of annual student assessment results during the last decade has greatly expanded the use of assessment data nationwide. Fueled by federally mandated Grade 3 to 8 testing in reading and math, together with the ability to track individual student achievement over time, federal and state policy mandates (e.g., No Child Left Behind [NCLB] Act of 2001, Race to the Top, Colorado SB163-08) have increased the stakes associated with student performance on state assessments. The intense interest in education accountability has led to a variety of methods to examine assessment outcomes for use in making judgments about education quality. Receiving particular interest are analyses of student academic growth. Growth models, as they are often called, have found favor as the preferred method for analyzing student achievement data for accountability purposes.

The appeal of student growth is apparent and certainly not without merit. Because learning is demonstrated by changes in student achievement from one point in time to another, an interest in the process of student learning is an interest in academic growth. Applied to accountability systems based upon large-scale annual assessment results, various methods have been put forward to examine student achievement over time, including value-added analyses, growth-to-standard analyses, growth-curve modeling, value tables, gain scores, and student growth percentiles. The variety of models reflects the various questions and purposes that growth analyses are designed to address. This plurality of purposes is frequently lost on stakeholders wanting to utilize growth analyses in their own specific context.

Complicating stakeholder efforts to grasp the scope and purpose of measuring student growth is a myriad of technical documentation that often supplies details on how growth calculations are performed but lacks broader explanations of the underlying rationale and assumptions: What the growth model is, what questions the growth model can (and cannot) answer, what test assumptions

(e.g., properties of the measurement scales) the growth model is based upon, and what valid inferences can be derived from the growth model results. The end result is often a system of data analysis that belies the promise that growth models bring for transforming educational accountability.

In this paper we unpack issues related to student growth by situating the discussion within three larger, intersecting topics: Measurement, longitudinal data analysis, and accountability. Discussions of student academic growth often traverse topics related to these issues without clearly explaining how student growth relates to each. We assert that by better articulating the relationship between growth and these three topics, education stakeholders across the spectrum will be better informed.

Measurement Issues

Fundamentally, an examination of student growth is an examination of student achievement over time. Without achievement measured for an individual at two or more points in time, it is impossible to consider achievement change, and hence, growth. This fact points to the central role that achievement testing plays in the analysis of growth. The quality and characteristics of the measurement scales on which student achievement is reported are fundamental for the types of growth analyses to be performed as well as the inferences made from the results. Growth analyses based upon impoverished measures of student achievement are themselves necessarily impoverished.

Before getting into a deep discussion of growth in achievement, it behooves us to first address the question: Achievement in what? Without an underlying construct against which achievement is referenced, any notion of growth or change has little meaning. For example, we wouldn't measure a child's height and, later, measure their weight and ask the question, How much growth occurred? Nor would we measure reading achievement and, later, mathematics achievement and ask a question about the amount of growth in achievement. Growth is thought to occur along some continuum associated with the relevant construct. The stringency with which the construct must be defined is debatable. At the most stringent, a time-invariant unidimensional construct might be considered necessary to reference change. Such constructs are difficult to establish in education where, for example, the mathematics test at Grade 8 that includes aspects of algebra may not measure the same dimension of mathematics as the Grade 3 mathematics test that focuses on basic arithmetic facts and operations. As a less stringent alternative, one might demand a construct labeled *reading*, indicating what is being examined over time. The more construct-specific the questions one wants to answer, the greater the stringency on the underlying construct.

With a construct established, there are numerous scales on which to report student achievement over time. The two most common scales used to report achievement are the ordinal performance levels upon which current criterion-referenced assessments are anchored (e.g., below basic, basic, proficient, and advanced) and scale scores. Scale scores themselves can be distinguished depending upon whether they are vertically linked or not. Vertically linked scales and their associated scores resemble the familiar measurement scales used to quantify height and weight. Such scales provide a cross-grade achievement continuum allowing for the comparison of student achievement at different points in time (e.g.,

different grade levels).[1] By contrast, nonvertical scales do not allow for the comparison of student scale scores from different grades. Both types of scales are common in large-scale assessment programs nationwide.

The ideal referent for measuring growth in student achievement is distance and its first derivative velocity in the physical sciences where, for example, a construct (e.g., a particle's position) is examined over time. The interval properties of scales used to measure, for example, an individual's height are, at best, approximated in educational measurement, making the analogy a fragile one. However, consideration of an ideal measurement scenario does help elucidate the phenomena we might ideally wish to consider, were such scales a possibility.

Measuring Growth With Achievement Scales: The Magnitude of Growth

Following the physical sciences, growth (or change) is conceptualized as a *magnitude* (i.e., a *distance*) governed by the relationship *distance = rate* x *time.* For example, distance travelled is often reported in miles and miles/hour is the metric for rate of change. Translating this example to the scores derived from large-scale annual assessments, the magnitude of growth, like distance, is conceptualized as an amount of change in the reporting metric from one year to the next, with rate of change indicating the magnitude/year.

If, instead of scale scores, performance standards are the desired achievement metric, student growth would be reported relative to changes in the performance level of a student, from one year to the next. For example, a student might make annual growth from *basic* to *proficient* on a state assessment, with growth indicated as the performance level change. Performance standards are a popular way of reporting achievement and they are required by NCLB. As the accountability metric of choice, they are a natural choice to anchor discussions of student performance, including growth.

An obvious limitation of using performance standards to report growth, however, is that the few levels associated with the performance standards (usually 3 or 4 levels) often contain a wide range of achievement and hence have the potential to mask substantial student growth. For example, a student might remain at the proficient level from one year to the next, even though significant growth might have occurred.[2]

Variability in the stringency of performance standards from state to state is also a reason for concern (Bandeira de Mello, Blankenship, & McLaughlin, 2009; Linn, 2007). Just as reports of student

[1] Scale scores derived from vertical scales resemble the interval scales used in the physical sciences though the interval properties of these scales are disputed (Ballou, 2008; Yen, 2007, 1986).

[2] Despite the fact that performance standards are not ideally suited to report growth at the *individual* student level does not necessarily imply that their use to quantify student growth at a *summary* level (e.g., student growth at the school level) suffers the same limitations. If descriptions and inferences about student growth are restricted to aggregate summaries, then performance standard-based growth might be suitable for the intended purpose. The goal is to match the preferred measurement scale with an appropriate analysis technique to the intended purpose in a valid manner.

achievement as a percentage of students scoring at or above proficient are contingent upon the state's performance standard stringency, performance standard-based analysis of student growth will also reflect each state's performance standards. As discussions about cross-state comparisons become more and more prominent, the necessity of having common metrics (e.g., standards) across states is apparent.

The granularity of scale scores overcomes one of the limitations of performance standards. However, utilizing scale scores to quantify the amount of growth creates other difficulties. Scale scores have the disadvantage that they are harder to understand by lay users of the results than performance standards. More significantly, the technical properties of vertically linked scores have been the subject of increasing scrutiny for the desired calculations and inferences they are often used for when examining growth (Ballou, 2008; Briggs & Betebenner, 2009).

A common approach to quantifying growth with scales scores requires the scales be vertically linked so that scale scores from one grade to the next can be compared and subtracted from one another. For example, the Colorado Student Assessment Program (CSAP) has vertically linked tests allowing individual student growth to be quantified in terms of scale score change (i.e., a gain score). Figure 1 depicts the change in scale scores between Grades 6 and 5 against the Grade 5 scale scores, together with a fit line to show average gains.

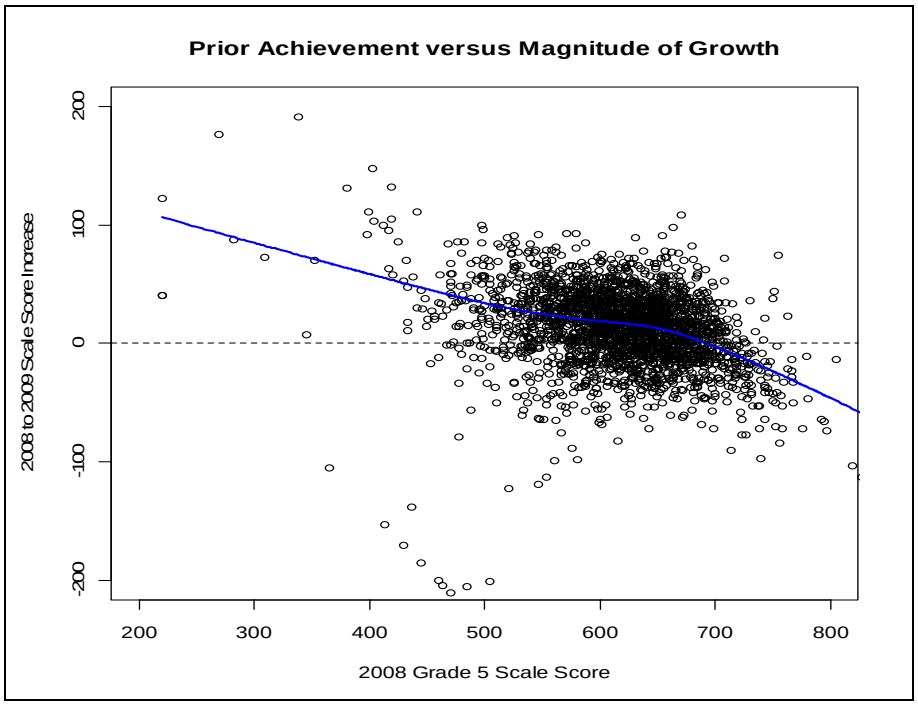

Figure 1. Magnitude of growth plotted against prior achievement, with a smoothed regression curve showing the well-known negative correlation between prior score and gain.

Close examination of Figure 1 indicates some of the difficulties associated with using scale score change across a vertically linked scale. Ceiling effects associated with commonly administered large-scale assessments often lead to negative scale score gains for very high achieving students. More importantly, comparison of gains in Figure 1 requires that the vertically linked scale have interval properties so that differences observed at different points of the scale can be compared. This is a nontrivial assumption and one that likely cannot be sustained (Ballou, 2008; Yen, 1986).[3]

Like gain scores, more sophisticated analysis techniques requiring cross-grade score comparability demonstrate qualities similar to gain scores, making their results difficult to support. Hierarchical, linear-model-based, growth-curve analyses (Singer & Willett, 2003) require a vertically linked scale where an outcome variable such as achievement is modeled as a function of time. In a recent study investigating the impact of scale on growth-curve analysis results, Briggs and Weeks (2009) constructed numerous, equally defensible vertically linked scales and show that growth-curve analyses can yield very different rates of growth, leading to different conclusions about a student's growth, particularly when interpreted relative to performance standards.

Even if vertically linked scales with interval properties can be produced for education assessment with properties closely approximating those in the physical sciences, it is not clear how these scales will make growth analysis and interpretation more transparent. Consider the following scenario where measurement scales are perfect:

A male child is measured at 2 and 3 years of age and is shown to have grown 4 inches. The magnitude of increase—4 inches—is a well understood quantity that any parent can grasp and calculate at home using a simple yardstick. However, parents leaving their pediatrician's office knowing only how much their child has grown would likely be wanting for more information. Parents are not interested in a magnitude of growth, but, instead, understanding that 4 inches in the context of similar children. Examining this height increase relative to the increases of similar children permits a diagnosis of how (ab)normal such an increase is (Betebenner, 2009, p. 44).

This simple example illustrates the necessity of context in understanding growth. Measurement alone (no matter how perfect) will not trivialize student growth calculations. We return to the theme of context later, in the discussion of growth norms and student growth percentiles (Betebenner 2008, 2009).

This example brings to light the critical issue of what questions stakeholders want growth analyses to answer. If the desire is to quantify whether growth is normal or abnormal (vis-à-vis growth norms), then a vertical scale is not necessary. However, it is important to note that some stakeholder questions require growth magnitudes. For example, if the desire is to determine whether low achieving students *grow as much* as high achieving students, then it is necessary to have an interval scale that can quantify growth along the entire achievement continuum. Similarly, determining whether students in later

[3] Computer adaptive testing shows promise in helping to remove ceiling effects associated with grade-level testing, yielding greater precision of student achievement scores, especially toward the ends of the achievement distribution.

grades *grow more* than in earlier grades requires a vertical and interval scale over which growth can be compared. In our opinion, it is doubtful that scales in education measurement will ever reach the level of technical quality necessary to answer such questions definitively.

If magnitude of growth is what one wishes to understand, arguably the most informative approach to describing the magnitude of growth involves a qualitative inventory of the progress a student makes over time. Learning progressions represent such an approach, meticulously defining a subject matter continuum that the student is traversing and itemizing the stages across which progress occurs. As such, a rich description of student progress is available at the individual level, indicating where the student was, where they are now, and the steps they have taken in the interim. It is unclear whether such rich descriptions can be *summarized* in a manner that makes them amenable to present accountability uses and mandates.

Measuring Growth With Achievement Scales: Magnitudes Versus Norms

Vertically linked scale scores provide a basis for comparing the magnitude of the growth made by students who start at different positions. Although the scales may fall short of the demands of a truly equal-interval scale, they often approximate an equal interval scale. Thus, it may be reasonable to conclude that a student who started with a scale score of, say, 300 and ended with a score of 330 (a gain of 30 points), grew more that a student who started with a scale score of 200 and ended with a scale score of 215 (a gain of 15 points). However, it is not possible to conclude that the second student grew twice as much as the first student or that the second student learned twice as much. Moreover, absent context, it is not possible to say if a gain of 15 or 30 points is large or small.

Given the difficulty of calculating growth magnitudes, together with the necessity for context in understanding them, what alternatives exist? The difficulty in interpreting scale scores to quantify growth strongly hints at trying to find a different metric to quantify change. One candidate borrows from pediatrics, where norm-referenced height and weight are common. Almost all parents have experience with the height and weight norms associated with infants. Growth norms anchor understanding, using this familiar percentile metric to motivate discussions about whether student progress is normal or abnormal.

Using growth norms does not allow for comparisons about magnitudes of growth because the percentile norms do not possess even approximate equal interval properties. However, growth norms are possible with or without a vertically-linked scale and provide a basis for making probabilistic comparisons about progress—quantifying whether one student's progress is more exemplary than another's. This provides information that is lacking in a simple analysis of scale score gains and is likely to be more helpful diagnostically.

The Colorado growth model provides the most developed approach today to using growth norms for large-scale state assessment. Each student with longitudinal data receives a *student growth percentile,* quantifying growth in a norm-referenced fashion. If it is known, for example, that a gain of 20 scale score points is at the 25th percentile for students who start with a scale score of 300, then it is reasonable to conclude that the gain of 20 scale score points is fairly small in comparison to the growth made by the student's peers. Without context, the 20-point magnitude of change has little relevance.

Measurement of student achievement provides various types of data that are amenable to the analysis of student growth. We believe that before embarking on an examination of student growth with available large-scale assessment data it is important to match the questions to be addressed with suitable achievement measures and longitudinal data analysis techniques. For example, vertically linked scales are necessary for some questions and not for others. However, in general, we find much of the rhetoric associated with the necessity of vertically linked scales for growth analyses to be overstated. Indeed, measurement scales can only do so much in addressing questions of growth. In particular, longitudinal data analysis techniques and inferences associated with accountability systems greatly inform the quality of growth analyses. In the next section we discuss the variety of statistical models available for growth analyses and their requirements concerning measurement scale as well as inferences that can be drawn from their results vis-à-vis accountability.

Longitudinal Data Analysis

There is no shortage of techniques currently being used to analyze longitudinal student assessment data. The diversity of methods is a reflection of the various purposes and associated questions that growth analyses are intended to address. Unfortunately, this variety of purposes and questions tends to get lost in discussions of growth, particularly in the policy arena. Clearly stated purposes and questions are often an afterthought in growth model development and adoption. Making these purposes and questions explicit will greatly benefit policy initiative like Race to the Top as well as ESEA re-authorization.

A critical first step in the statistical analysis of any data is in specifying the purpose of the analysis. What questions will the analysis address and, at best, answer? Tukey's (1962) maxim that it is better to have an approximate answer to the right question than an exact answer to the wrong question speaks to the prominent role of the question in data analysis. Longitudinal data analysis is no exception.

Potential questions addressed by student growth analyses span a wide range and are often motivated by external accountability mandates. Simultaneously, as addressed in the previous section, the questions growth analyses can answer are constrained by the achievement measures on which they are constructed. Currently, state departments of education are struggling to engineer growth analysis techniques that meet the often conflicting demands of accountability mandates and stakeholder expectations while simultaneously taking account of the properties of the assessment instruments used.

What are the questions of interest to stakeholders? In a survey conducted by Yen (2007), parents, teachers, and administrators were asked what questions related to student growth were of most interest to them:

Parent questions

- Did my child make a year's worth of progress in a year?
- Is my child growing appropriately toward meeting state standards?
- Is my child growing as much in math as reading?

- Did my child grow as much this year as last year?

Teacher questions:

- Did my students make a year's worth of progress in a year?
- Did my students grow appropriately toward meeting state standards?
- How close are my students to becoming proficient?
- Are there students with unusually low growth who need special attention?

Administrator questions

- Did the students in our district/school make a year's worth of progress in all content areas?
- Are our students growing appropriately toward meeting state standards?
- Does this school or program show as much growth as that one?
- Can I measure student growth even for students who do not change proficiency categories?
- Can I pool together results from different grades to draw summary conclusions?

Recently, Race to the Top has highlighted another set of questions that align with accountability initiatives and are more attributional in nature. Namely, value-added-type questions associated with the contributions of teachers and principals to student achievement are prominent. In order to receive funding in line with Race to the Top, answers to these questions must be used in judgments about educator quality. In what follows we consider some common types of growth models and the questions and purposes they address, together with any properties of assessments that they rely upon.

Value-Added Models

A primary use of growth analyses over the last decade has been to parse the amount of student growth that can be attributed to the school or teacher (Ballou, Sanders, & Wright, 2004; Braun, 2005; Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004). Such analyses, often called value-added analyses, attempt to estimate the teacher or school contribution to student achievement. This contribution, called the *teacher or school effect*, purports to quantify the impact on achievement that this school or teacher would have, on average, upon similar students assigned to them for instruction.[4]

The attraction of such analyses in an era of accountability is clear. For example, recent guidance for Race to the Top applications required states to be:

> Differentiating teacher and principal effectiveness based on performance:….The extent to which the State, in collaboration with its

[4] In recent Race to the Top guidance, the term *principal effect* has been substituted for *school effect*.

> participating LEAs, has a high quality plan and ambitious yet achievable annual targets to (a) Determine an approach to measuring student growth (as defined in this notice); (b) employ rigorous, transparent, and equitable processes for differentiating the effectiveness of teachers and principals using multiple rating categories that take into account data on student growth (as defined in this notice) as a significant factor; (c) provide to each teacher and principal his or her own data and rating; and (d) use this information when making decisions…. (Race to the Top, p. 37809).

Such directives are consistent with federal and state accountability initiatives over the past decade linking student assessment outcomes to judgments of education quality (e.g., NCLB and Colorado SB 163-08).

The terms *value-added analysis* or *value-added assessment* suggest a particular type of analysis or assessment. Instead, value-added refers primarily to the manner in which the results from an analysis of the assessment data are used to make value-added (i.e., causal) attributions about responsibility for the outcomes. Value-added analyses begin with annual student assessment data and by employing different statistical controls (which can vary depending upon the type of analysis performed) they attempt to tease out the contribution associated with the teacher, school, and or principal on a student's achievement.

There is a growing body of literature that scrutinizes the validity of the value-added procedures and the estimates they produce (Braun, 2005; Rubin et al., 2004). At issue is whether observational data can be used to infer the effectiveness of a given teacher or school and, more fundamentally, what the terms *teacher effect* or *school effect* actually mean (Raudenbush, 2004). Rubin et al. (2004) suggested that such effects might be useful as descriptive measures in that they potentially provide actionable data that can be used by stakeholders to improve the quality of education.

The most prominent example of value-added analyses are those associated with the Education Value-Added Assessment System (EVAAS) system first implemented in Tennessee and currently in use in Ohio, Pennsylvania and school districts throughout the United States. The EVAAS system employs a multivariate, mixed-effects analysis to produce teacher or school effects (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Sanders, Saxton, & Horn, 1997). In addition to EVAAS there are a number of other value-added analysis techniques being used for both research and accountability purposes throughout the United States. From a measurement perspective, the majority of techniques employed for value-added purposes do not require a vertically linked scale but do require a scale with interval properties (Ballou, 2008; Briggs & Betebenner, 2009).

Value-added analyses return norm-referenced effectiveness quantities indicating whether a teacher, school, and/or principal is significantly more or less effective than the norm-groups average (e.g., a district or state). Figure 2 depicts 2008 Tennessee Value-Added Assessment System (TVAAS) elementary school effects against school poverty. Note that the effects are centered about the horizontal zero, indicating average effectiveness. Like all norm-referenced quantities, value-added effects don't implicitly include criteria stipulating what is *enough* effectiveness or criteria distinguishing *good* levels of

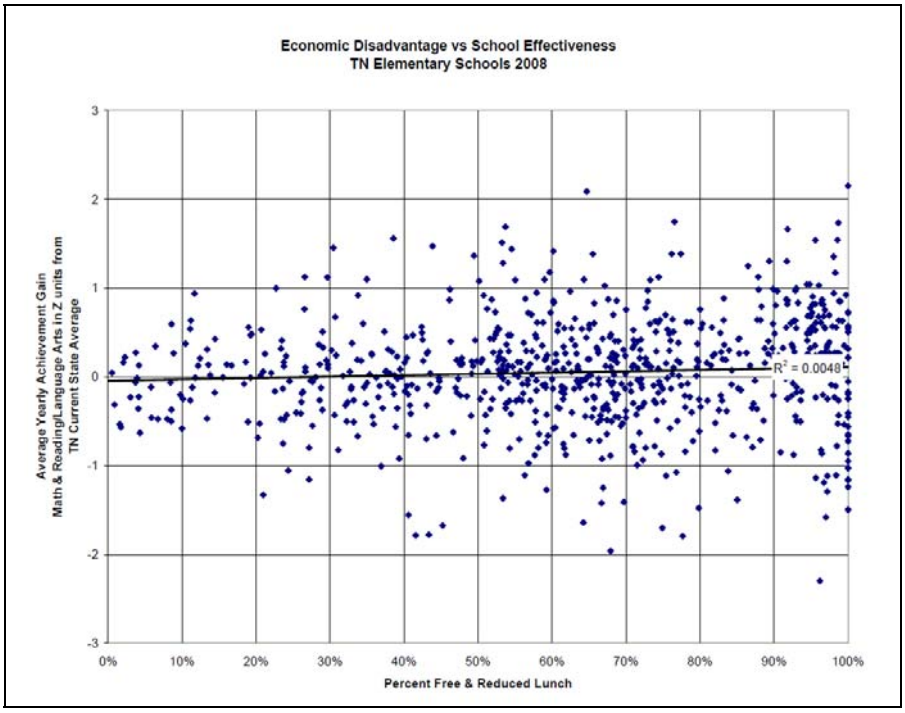

Figure 2. Elementary school poverty versus value-added school effects, based upon 2008 TVAAS data from the Education Consumers Foundation (2008).

effectiveness from *bad*. For example, considering the universal proficiency mandates associated with NCLB, what must the value-added effects be for schools in order to accomplish this policy goal? Value-added models do not seamlessly fit within a criterion-referenced assessment and accountability system currently in existence.

The limited ability of value-added models to easily accommodate criterion-referenced ends again points to the critical nature of identifying purposes and questions in the process of establishing a growth model. Value-added analyses are designed to answer some questions. Stakeholders must decide whether those questions address stakeholder interests and policy mandates.

The interest in incorporating growth measures into criterion-referenced accountability led to the development of criterion-referenced growth-to-standard models that, as their name suggests, stipulate adequate progress in light of students reaching desired achievement outcomes (Betebenner, 2009). These models establish individual adequacy expectations around policy mandated achievement outcomes instead of statistical expectation.

Growth to Standards

Given the popularity of using student progress to make judgments about education quality, it is not surprising that the November 2005 announcement by Secretary of Education Spellings of the Growth Model Pilot Program (GMPP) permitting states to use growth model results as a means for compliance with NCLB achievement mandates was met with great enthusiasm by states (Spellings, 2005). As

guidance to states applying for the GMPP, the U.S. Department of Education stated explicitly that the universal proficiency mandate of NCLB would not be compromised and that growth models would be held to the same exacting standard as approved status models. Yen (2009) provided a comprehensive account of all of the 15 accepted state growth models.

Of the approved models, a majority maintain compliance by judging growth based upon future student achievement, judging whether a student is on track to reach proficiency or some other future achievement outcome. Referred to as the growth-to-standard approach, these criterion-referenced growth models designate whether a student is *on track to being proficient* and invoke this designation, usually in conjunction with other status measures, as evidence of school quality.[5] The models utilize differing means of determining whether students are on track to reach the designated targets, including the calculation of growth trajectories, projected future achievement, and growth norms. Operationalizing growth relative to future achievement targets represents a departure from more familiar student growth models, including the widely discussed value-added models. More broadly, however, this departure represents an attempt to anchor growth to policy-mandated criteria, as opposed to the more familiar norm-referenced criteria embedded within value-added models (Betebenner, 2009).

Although the GMPP is a step in the right direction, it has had little practical effect on the classification of schools as making or failing to make AYP, because of a severe constraint that is imposed on states using a growth model. The constraint is that students who are not already proficient must have growth that is sufficient to bring them to the proficient level or above within three years. In studies by Dunn (2007, 2009), results from status and growth-to-standard models were compared to status and various other growth models. Her findings indicate that the NCLB-approved GMPP models classify schools very similarly to status models. Students who are far below the proficient level have little chance of growing at a rate that will bring them up to that level within three years. Consequently, a school where the majority of students are far below the proficient level gains little if anything from the growth model pilot. By and large, the same schools that were failing to make AYP under the status model also fail to make AYP under the growth model pilot. Due to the their close alignment with status—using growth to estimate future achievement—growth-to-standard models represent an impoverished view of growth and serve, more generally, to limit the concept of growth as it relates to student achievement.

Descriptive Growth Models: Growth Norms

Often engineered to satisfy accountability mandates, growth models, particularly value-added models, are purposed with finding causal effects associated with, for example, teachers or schools. The use of statistical analyses in such endeavors does not represent the only way in which longitudinal data can be analyzed. Stepping back and looking at how data is analyzed more broadly, there are three general uses associated with statistical models (Berk, 2004):

[5] These models are referred to in the literature by various other names including the hybrid success model (Kingsbury, Olson, McCahon, & McCall, 2004) and the REACH value-added model (Doran & Izumi, 2004).

1. **Description**—An account of the data. Model is true to the extent that it is useful and quality judged based upon craftsmanship (de Leeuw, 2004).

2. **Inference**—Sample to Population. Model is true to the extent that the assumed chance process reflects reality.

3. **Causality**—*A* causes *B* to happen. Model is true to the extent that plausible causal theory exists and design criteria are met.

In practice, models are rarely descriptive, despite minimal requirements associated with their specification. Instead, statistical models usually push toward inferential and/or causal uses that are often difficult if not impossible to justify given the nature of the data or the causal hypothesis specified. The influence of accountability on the analysis of growth has pushed the statistical models of growth to focus almost exclusively on issues of causality to the exclusion of other uses such as description.

At their base, growth measures are descriptive. With the focus on accountability and attributions of responsibility, growth measures have been crafted specifically for use in quantifying value-added effects. It is important to note, however, that growth measures do not have to address only causal ends. Good descriptive measures are interpretable, informative, and capable of multiple uses. Descriptive growth models like the student growth percentiles used as part of the Colorado growth model results are used for multiple purposes, including the investigation of schools that have students demonstrating disproportionately high or low growth.

Seen as a part of a larger research program directed toward identifying causes in complex social systems, descriptive growth measures such as growth norms and student growth percentiles

- are helpful in spotting provocative associations,
- are a part of advocacy/informative discussions (e.g., growth gaps by ethnicity), and
- inform policy goals and initiatives.

Given the wide variety of models presented, we must reemphasize that the quality of a growth model cannot be determined from the model itself. The questions and purposes the growth model is asked to address must be at the forefront in any determination of model quality. In this section, we have delineated the types of questions, from descriptive to causal and from norm- to criterion-referenced that growth models are today tasked with addressing. Because longitudinal data analysis techniques do not answer all questions equally well, stakeholders must match the model with stakeholder questions and purpose, to give the growth model the best chance of fulfilling its promise to transform the way in which education is viewed.

Accountability

Calls for the incorporation of student growth into educational accountability represent the latest chapter in a two-decade expansion of educational oversight based upon large-scale assessment results. Beginning in the 1990s, a number of states introduced test-based accountability systems. Encouraged at

the federal level by the Goals 2000 Act of 1994 and in the Improving America's Schools Act (IASA) of 1994 (which reauthorized the Elementary and Secondary Act [ESEA] of 1965), standards-based accountability systems came to the fore. Despite their prominence, however, enforcement of IASA requirements was fairly limited.

In the first decade of the 2000s, test-based accountability plays an increasingly prominent role due primarily to the latest reauthorization of the ESEA—NCLB. The test-based accountability requirements are more sweeping and explicit in NCLB than they were in earlier ESEA reauthorizations. Moreover, the U.S. Department of Education has been much more aggressive in enforcing the state accountability requirements of NCLB than it was with IASA. As a consequence, states that had not already done so were forced to introduce tests in mathematics and reading or English language arts at Grades 3 through 8 plus one grade in high school as well as science tests in at least one grade in each of three levels: elementary, middle, and high school. They have also had to either replace their state system of accountability with the NCLB accountability system or run two separate accountability systems, which sometimes results in conflicting categorizations of schools.

Status-Based Accountability

At the onset of NCLB, prior to the GMPP, all state-approved NCLB accountability systems utilized status-based approaches to judging education quality. States were required to establish performance standards that define proficient performance in mathematics and reading or English language arts at each grade level. They also had to define intermediate, system-level improvement targets, called annual measurable objectives (AMOs), which establishes universal proficiency for all students by 2014 as the accountability mandate nationwide. Schools must meet or exceed the AMO in mathematics and in reading or English language arts for all students and for each of several reporting subgroups for which there are sufficient students in the school for disaggregated reporting, in order to be identified as having made adequate yearly progress (AYP).

NCLB requires states to use student assessment results to determine whether or not schools make AYP each year, and schools that fail to meet AYP for two or more years in a row are subject to *corrective actions*. Schools that continue to fail to make AYP are subject to increasingly severe sanctions each successive year after falling into the needs improvement category. Schools that fail to make AYP are implicitly judged to be of lower quality or less effective than schools that make AYP. The validity of this school quality interpretation is problematic, since there are many reasons other than differences in school quality or relative effectiveness that may result in one school making AYP while another does not.

Achievement in any given year may fall short of the AMO because the school is ineffective. There are, however, a host of other reasons besides ineffective instruction that can lead to low achievement. The students may have had low achievement in previous years and despite substantial growth in the year in question, may still fall short of the AMO. A school that makes AYP may have students who started the year with relatively high achievement as the result of favorable home conditions and support whereas a school that fails to make AYP may do so because its students start the year with low achievement as the result of unfavorable family conditions and educational support in prior years. As is illustrated in Figure 3, there is a substantial negative relationship between poverty as indicated by the percentage of

students in a school on free or reduced lunch and the percentage of students in a school who are proficient. The relationship is greatly reduced by the use of growth in achievement rather than achievement status (Figure 4).

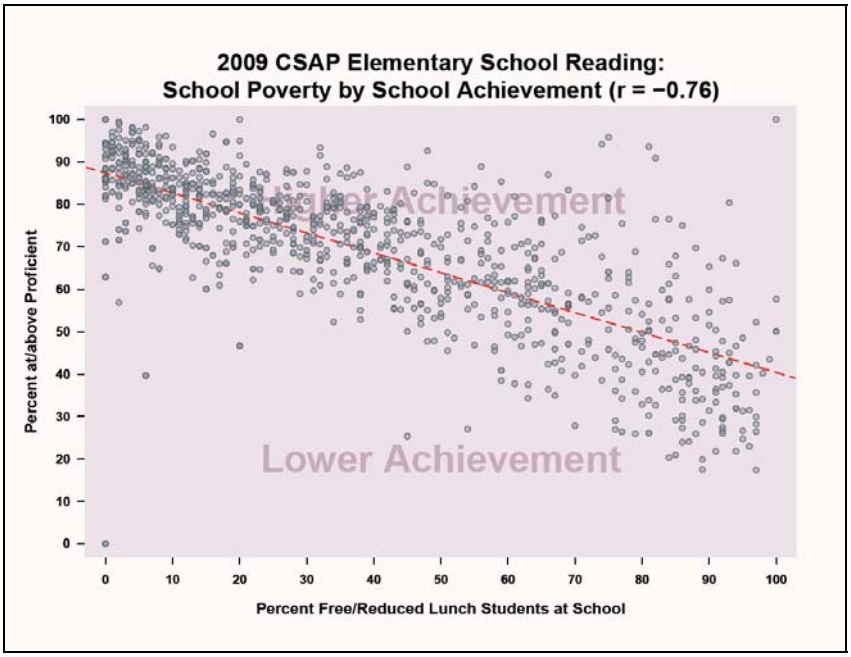

Figure 3. 2009 Colorado elementary school poverty versus achievement (percentage of students at/above proficient) in reading.

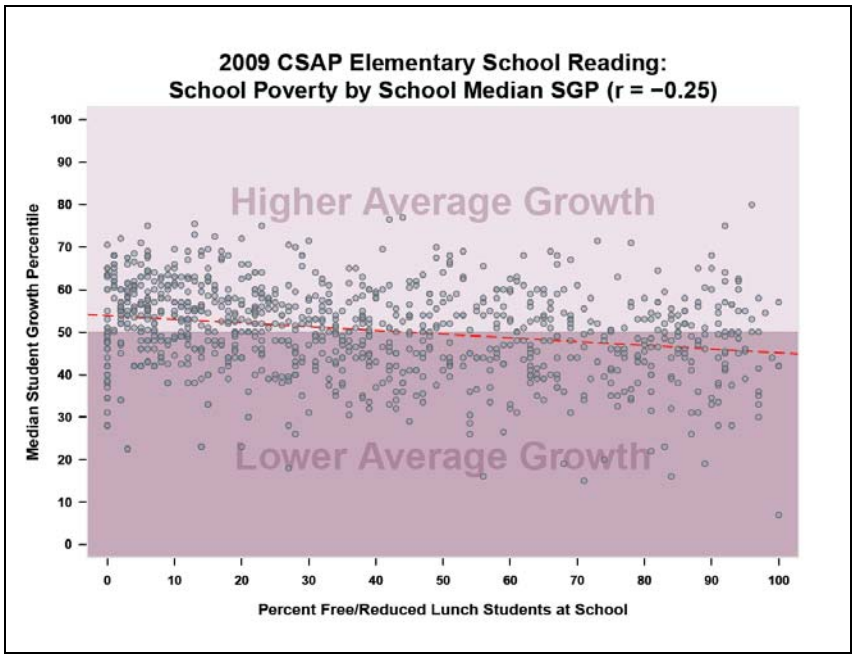

Figure 4. 2009 Colorado elementary school poverty versus growth (median student growth percentile) in reading.

Growth and Status

Growth in achievement and achievement status are both important considerations in judgments of school quality and school accountability. Growth is important because it provides a more direct measure of student learning than status does. Growth is also important because it provides a more equitable basis for comparing schools that serve student populations that differ in family backgrounds and in prior achievement than a simple comparison in terms of end-of-year achievement. On the other hand, it is also important to consider the level of current achievement to avoid a dual system that sets different standards for different groups of students. Thus, it is better to base accountability on a combination of growth and status rather than on either alone.

There are several ways in which both growth and status can be taken into account in a school accountability system. A minimum level can be established for current performance and distinctions can then be made using a growth measure. Alternatively, accountability can be based on a point system where points are earned for both rate of growth and the level of current achievement. A third approach currently being used in Colorado utilizes the statewide growth norms to synthesize the growth norms with a growth-to-standard approach. Combining norm and criterion-referenced growth in this way allows the state to balance policy demands for what should be with the empirically based growth norms to promote ambitious yet reasonable achievement outcomes for all students.

For any approach that takes into account both status and growth, it is important to set ambitious goals, but ones that are reasonable for schools to reach given sufficient effort. One way of determining whether goals are reasonable is to base them on normative information obtained from past experience with high performing schools. It is also possible to use a combination of norm-referenced and criterion-referenced performance and growth goals.

Accountability System Objectives and Consequences

Accountability systems are intended to improve education. More specifically, they are intended to (a) increase student achievement, (b) reduce achievement gaps, and (c) increase efficiencies. Accountability systems are expected to accomplish these objectives by imposing sanctions and rewards based on results from large-scale assessment outcomes.

The objectives of accountability systems are laudable. It is important, however, to evaluate the degree to which these laudable objectives are achieved. It is also important to evaluate unintended negative side effects that may be associated with the accountability system and the degree to which variations in the accountability system are likely to have more positive consequences and less severe negative effects. In other words, it is critical to evaluate the validity of the accountability system. An approach to judging the validity of an accountability system that has been developed by Henry Braun is to consider its utility: “Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (access, quality, equity, efficiency) within an educational system, without causing undue deterioration with respect to other goals” (Braun, 2008)

The U.S. Department of Education has put in place a system of peer reviews that are intended to ensure that state standards and assessments meet the requirements of NCLB. In connection with the peer review process, the Department has published *Peer Review Guidance* (U.S. Department of Education, 2004), which provides states and reviewers with fairly detailed specifications of the evidence that states are expected to accumulate and peer reviewers are expected to evaluate. *Peer Review Guidance* requires states to provide several types of validity evidence. Included among the types of evidence is a consideration of the consequences of assessment uses for purposes of NCLB accountability: "States must attend not only to the intended effects, but to unintended effects" (U.S. Department of Education, 2004, p. 33).

It is not easy to obtain evidence that the assessments and the accountability system have utility, maximize the intended effects, and minimize the unintended negative effects to satisfy the demands of *Peer Review Guidance* or, in Braun's terms, the demands that they have systemic validity (Braun, 2008). Without such evidence, however, the use of assessment and accountability system results to sanction schools is hard to justify. There are a variety of techniques that can provide data to be used as evidence relevant to judging the plausibility that assessment results have a specified set of effects. Questionnaires, interviews, observations, focus groups, the collection of data of record (e.g., course-taking patterns, graduation, and dropout rates), and the collection of instructional artifacts (e.g., student assignments, and classroom tests) can be used to collect the needed evidence. The collection, reporting, and interpretation of such evidence needs to be given greater priority than it has enjoyed in the past—the data doesn't speak for itself.

Data Use: Descriptive Versus Causal Interpretations

Accountability systems are often thought of and treated as if they produce causal information. Schools that fail to meet targets established as part of the accountability system may be subject to sanctions of various types of corrective actions regardless of any extenuating circumstances or evidence that is inconsistent with the internal results of the accountability system. Some descriptions such as value-added effects have strong causal connotations that go beyond the more descriptive claims that authors such as Raudenbush (2004) and Rubin et al. (2004) argued are justified for value-added models. Accountability system results do not have to be used to make causal interpretations to be useful, however:

> Accountability system results can have value without making causal inferences about school quality, solely from the results of student achievement measures and demographic characteristics. Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics are potentially of considerable value. Rather than using the results of the accountability system as the sole determiner of sanctions for schools, they could be used to flag schools that need more intensive investigation to reach sound conclusions about needed improvements or judgments about quality (Linn, 2008, p. 21).

Summary and Next Steps

There is a tremendous interest in measuring growth and in using growth as part of accountability systems for both teachers and schools. The expansion of achievement testing in all states, together with the establishment of longitudinal data bases containing student records with achievement test results over two or more years has made it feasible to use the data for measuring student growth. There are a variety of approaches to the analysis of growth The different approaches answer different questions and the interpretations that are appropriate for one approach or model may not be appropriate for a different approach.

There are at least three intersecting considerations that are critical to the understanding and selection of a growth model. These are measurement considerations, longitudinal data analyses considerations, and accountability considerations. Accountability considerations, such as high-stakes attributions of responsibility or universal proficiency mandates, tend to impose themselves upon measurement and longitudinal data analysis considerations, often to the detriment of constructing measures based upon student growth (and of using them sensibly.) To remedy this, stakeholders must strike a delicate balance between measurement, longitudinal data analysis, and accountability considerations.

As states rapidly move forward in the coming years and mine student growth from their longitudinal data systems, it is important to not ignore hard learned lessons. Large-scale assessment data represents an important, yet incomplete, set of evidence from which to judge school and educator quality and inform practice. The larger task of marshaling evidence to motivate, evaluate, and inform our educational practices requires painstaking detective work and attention to detail. Growth data is just one data point in a larger data ecosystem that is being increasingly tapped in the hopes of improving education. In addition to making best use of currently available large scale assessment data, stakeholders should simultaneously analyze and report other evidence (e.g., interim assessment data) to supplement the incomplete picture supplied by large scale assessment data. State longitudinal data systems that are rapidly making such data available and using it to triangulate our judgments about education quality will only improve the quality of our judgments.

The drive toward increasing data availability and quality are positive steps but subject to the many well-understood and often-recited lessons about the use of large-scale assessment data for high-stakes purposes (see, for example, Linn, 2000; Mintrop & Sunderman, 2009). They include: Narrowed curriculum, gaming the system (e.g., focus on bubble students), overemphasis on test preparation, as well as the frustration and demoralization of those most involved with school improvement efforts. Education is not unique in this experience, as these adverse consequences have been known for quite some time and are codified as Campbell's law (Campbell, 1976):

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. 49)

Given what we know from experience, how do we move forward and avoid repeating the same mistakes once again?

It is helpful to step back from the numerous technical considerations and recognize some fundamental truths at the heart of any student growth analysis endeavor. First, growth analyses produce more data for education agencies already drowning in data. Moreover, these data do not speak for themselves. Once produced, student growth data can tell hundreds of stories. If systemic validity involves turning growth model data into useful information, then incorporating growth into accountability in a meaningful way requires much more than just a growth model. Based upon our experience consulting on and implementing growth models at the state and district level, what is necessary is a detailed theory of action delineating how student growth data leads to increases in education efficacy (Braun, 2008). This theory of action provides a template specifying what data is necessary and how that data will be used that in turn leads to actions with the desired outcomes.

Indeed, of the three issues (measurement, longitudinal, data analysis, and accountability) that currently impact growth modeling, the one of greatest significance is accountability. Though measurement and longitudinal data analysis issues are numerous, it is accountability that is pervasive and carries the greatest weight. Investigations of growth models are not academic exercises whose validity is established solely by technical considerations. Their central role in accountability discussions mandates the fundamental role of systemic validity in judging their quality. Box (1987) emphasized this point more colloquially when he stated, "All models are wrong, but some are useful" (p. 424).

This point circles back to one made in the discussion of longitudinal data analysis where the primary task is to establish questions followed by a data analysis technique to answer those question. Often, technical considerations obscure more pragmatic concerns, such as questions to be answered, stakeholder interests, and model usefulness. Considerations about data (e.g., student growth data) should center at least as much upon data use as on data quality. Placing greater emphasis on pragmatic concerns leads model development toward the use of cases that maximize potential utility, and, hence, systematic validity.

Emphasizing data use begs the question: Data use by whom? Which stakeholders (e.g., policy makers, administrators, teachers, and/or parents) should use data, which data should they use, and how do we envision them using it? Answering this question involves a thorough explication of the theory of action associated with increases in education efficacy. Breiter and Light (2006) reported on the difficulties of data use and pointed out several impediments to effective data use, particularly in situations involving practitioners. Much greater care must be taken to get the right data, to the right people, at the right time, and in the right format. Turning data into information and ultimately into knowledge requires concerted effort that involves striking an ideal balance between data quality, data availability, and data use.

References

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007* (NCES 2010-456). Washington, DC: National Center for Education Statistics.

Ballou, D. (2008, April). *Test scaling and value-added measurement*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Taylor & Francis.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Education Measurement: Issues and Practice, 28*(4), 42–51.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surface.* New York, NY: Wiley.

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

Braun, H. (2008, September,). *Vicissitudes of the validators*. Paper presented at the Reidy Interactive Lecture Series, Portsmouth, NH.

Breiter, A., & Light, D (2006). Data for school improvement: Factors for designing effective information systems to support decision-making in schools. *Educational Technology & Society*, *9*(3), 206–217.

Briggs, D. C., & Betebenner, D. W. (2009, April). *Is growth in student achievement scale dependent?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Campbell, D. T. (1976), *Assessing the impact of planned social change* (Occasional Paper Series, 8). Hanover, NH: Dartmouth College. The Public Affairs Center.

de Leeuw, J. (2004). Preface. In R. A. Berk (Author), *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Doran, H. C., & Izumi, L. (2004). *Putting education to the test: A value added model for the state of California* (Tech. Rep.). San Francisco, CA: Pacific Research Institute.

Dunn, J. L. (2007, September). *When does a "growth model" act the same as a* status model*: Lessons learned from some empirical growth model comparisons*. Paper presented at the Systems and Reporting conference, SCASS, Nashua, NH.

Dunn, J. L., & Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: coordinating measurement and accountability. *Education Measurement: Issues and Practice, 28*(4), 27–41

Education Consumers Foundation. (2008). *Economic disadvantage vs. school effectiveness. TN Elementary Schools 2008.* Retrieved from http://www.educationconsumers.org/tnproject/poverty_vs_effectiveness_2008.pdf

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10.

Goals 2000: Educate America Act of 1994, Pub. L. No. 103-227.

Improving America's Schools Act of 1994, Pub. L. No. 103-382.

Race to the Top, 74 Fed. Reg. 144 (July 29, 2009).

Kingsbury, G. G., Olson, A., McCahon, D., & McCall, M. (2004, July). *Adequate yearly progress using the Hybrid Success Model: A suggested improvement to No Child Left Behind* (Tech. Rep.). Portland, OR: Northwest Evaluation Association.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher 29*(2), 4–16.

Linn, R. L. (2007). Performance standards: What is proficient performance? In C. E. Sleeter (Ed.), *Facing accountability in education: democracy and equity at risk* (pp. 112–131). New York, NY: Teachers College Press.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3–24). New York, NY: Routledge.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

Mintrop, H., & Sunderman, G. L. (2009). *Why high stakes accountability sounds good but doesn't work—and why we keep on doing it anyway.* Los Angeles, CA: The Civil Rights Project/Proyecto Derechos Civiles at UCLA.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 § 115 Stat. 1425.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 120–129.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (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 evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University.

Spellings, M. (2005, November). *Secretary Spellings announces growth model pilot* [Press Release]. Retrieved from the U.S. Department of Education website: http://www.ed.gov/news/pressreleases/2005/11/1182005.html

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics 33*(1), 1–67.

U.S. Department of Education. (2004, April 28,). *No Child Left Behind. Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left*

Behind Act of 2001. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York, NY: Springer.

Yen, W. M. (2009). *Growth models for the NCLB growth model pilot*. Princeton, NJ: ETS.