EDUCATIONAL MEASUREMENT:
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2009, Vol. 28, No. 4, pp. 42–51

Norm- and Criterion-Referenced Student Growth

D. Betebenner, *National Center for the Improvement of Educational Assessment*

Annual student achievement data derived from state assessment programs have led to widespread enthusiasm for statistical models suitable for longitudinal analysis. The current policy environment's adherence to high stakes accountability vis-à-vis No Child Left Behind (NCLB)'s universal proficiency mandate has fostered an impoverished view of what an examination of student growth can provide. To address this, student growth percentiles are introduced supplying a normative description of growth capable of accommodating criterion-referenced aims like those embedded within NCLB and, more importantly, extending possibilities for descriptive data use beyond the current high stakes paradigm.

Keywords: growth, student growth percentiles, value-added models, normative student growth, criterion-referenced student growth

Accountability systems constructed according to federal adequate yearly progress (AYP) requirements currently rely upon annual "snapshots" of student achievement to make judgments about school quality. Since their adoption, such *status measures* have been the focus of persistent criticism (Linn, 2003; Linn, Baker, & Betebenner, 2002). Though appropriate for making judgments about the achievement level of students at a school for a given year, they are inappropriate for judgments about educational *effectiveness*. In this regard, status measures are blind to the possibility of low-achieving students attending effective schools. It is this possibility that has led some critics of No Child Left Behind (NCLB) to label its accountability provisions as unfair and misguided and to demand the use of growth analyses as a better means of auditing school quality.

A fundamental premise associated with using student growth for school accountability is that "good" schools bring about student growth in excess of that found at "bad" schools. Students attending such schools—commonly referred to as highly effective/ineffective schools—tend to demonstrate extraordinary growth that is causally attributed to the school or teachers instructing the students. The inherent believability of this premise is at the heart of current enthusiasm to incorporate growth into accountability systems. It is not surprising that the November 2005 announcement by Secretary of Education Spellings for 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).

Consistent with current accountability systems that hold schools responsible for the assessment outcomes of their students, the primary thrust of growth analyses over the last decade has been to determine, using sophisticated statistical techniques, the amount of student progress/growth attributable 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/school contribution to student achievement. This contribution, called the *school effect* or *teacher 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. Clearly, such analyses lend themselves to accountability systems that hold schools or teachers responsible for student achievement.

Despite their utility in high stakes accountability decisions, the causal claims of teacher/school effectiveness addressed by value-added models (VAM) often fail to address questions of primary interest to education stakeholders. For example, VAM analyses generally ignore a fundamental interest of stakeholders regarding student growth: How much growth did a student make? The disconnect reflects a mismatch between the questions of interest and the statistical model employed. In this direction, Harris (2007) distinguishes value-added for program evaluation (VAM-P) versus value-added for accountability (VAM-A). More broadly, the current climate of high-stakes test-based accountability combined with the emphasis of value-added toward school and teacher effects has skewed discussions about growth models toward causal claims at the expense of description. Research (Yen, 2007) and personal experience suggest stakeholders appear more interested in the reverse: description first that can be used secondarily as part of causal fact finding.

In a survey conducted by Yen (2007), supported by the author's

D. Betebenner is a Senior Associate at the National Center for the Improvemernt of Educational Assessment, PO Box 351, Dover, NH 03821-0351; dbetebenner@nciea.org.

own experience working with state departments of education to implement growth models, parents, teacher, and administrators were asked what "growth" questions were most of 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/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?

As Yen remarks, all these questions rest upon a desire to understand whether observed student progress is "reasonable or appropriate" (Yen, 2007, p. 281). More broadly, the questions seek a description rather than a parsing of responsibility for student growth. Ultimately, questions may turn to who/what is responsible. However, as indicated by this list of questions, they are not the starting point for most stakeholders.

Borrowing concepts from pediatrics used to describe infant/child weight and height, this paper introduces *student growth percentiles* (Betebenner, 2008). These individual reference percentiles sidestep many of the thorny questions of causal attribution and instead provide descriptions of student growth that have the ability to inform discussions about assessment outcomes and their relation to education quality. A purpose in doing so is to provide an alternative to punitive accountability systems geared toward assigning blame for success/failure (i.e., establishing the cause) toward *descriptive* (Linn, 2008) or *regulatory* (Edley, 2006) approaches to accountability.

In the introductory chapter to *The Future of Test Based Educational Accountability*, Linn (2008) describes such a descriptive approach:

> 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. (p. 21)

Christopher Edley (2006), in his invited presidential address at the 2006 AERA conference, expresses similar sentiments:

> This is the difference between a retrospective question of identifying fault as opposed to a prospective strategy to engineer some corrective measure, almost independent of considering whether there was blame-worthiness. And to move away from the blame-worthiness paradigm toward something that is more regulatory in nature where one might seize upon disparities or circumstances that are for some reason deemed unacceptable and engineer the interventions needed to bring about the necessary change. . . . It's the no-fault gap closing strategy in which the effort is to build a consensus about a vision of an improved society rather than figure out where's the person . . . we want to pillory.

As Linn (2008) notes, such an accountability system would represent a profound change from current systems. An essential first step toward such a change is the creation of appropriate and compelling descriptive measures on which to base the system. The following overview of student growth percentiles within the context of normative and standards-based growth is a first step in that direction.

Status and Growth

The impact of NCLB upon research connecting large-scale assessment outcomes with school quality has been profound. Current discussions often differentiate between accountability models/systems based upon status (i.e., achievement) and those based upon growth (Braun, 2005; Carlson, 2001; Hill, 2002; Hill, & DePascale, 2002; Linn et al., 2002). The rigid semantical distinction between status and growth models obscures their common foundation: namely, to understand student achievement via assessment outcomes. What considerations, if any, are necessary to understand students' level of achievement? The fundamental distinction between status and growth models is whether or not additional considerations—specifically prior achievement—should be taken into account to understand current achievement.

Status models, as their name implies, qualify student performance solely in terms of the current achievement (i.e., status) of the student. As such, status models are *unconditional achievement* models, examining student performance at a point in time with no conditioning variables. The output from such models within the criterion-referenced assessment systems found in all states is usually a dichotomous qualification (proficient/not proficient) of achievement for each student based upon the state's performance standards. As the basis for an accountability system with universal, rigorous achievement standards, such models are extremely demanding, requiring, without condition, an acceptable level of achievement from all students.

A natural extension to the basic characterization of achievement provided by status models is to qualify current achievement in terms of prior achievement. That is, what can be said of a student's current achievement level given their prior achievement? Conditional achievement models, or growth models, evaluate student progress based upon a longitudinal record of student achievement.[1] Figure 1 depicts the distinction of growth versus status as a difference between whether or not achievement is examined unconditionally or conditionally.

Situated between growth and status in Figure 1 are the projected

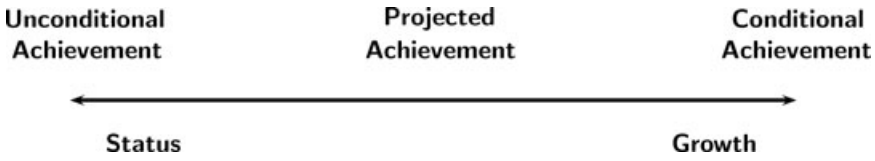

FIGURE 1. Unconditional, projected, and conditional achievement and their relationship to status and growth.

achievement (i.e., growth-to-standard) models, a popular use of growth accommodating NCLB achievement mandates (Auty & Group, 2008). Using a variety of statistical procedures, the models predict the future achievement of the student (usually up to 3 years) and the results are used in accountability systems to give schools credit for students on track to being proficient. That is, growth-to-standard models use a prediction of future achievement to make a determination about whether the student's growth is adequate. As such, growth-to-standard models are a criterion-referenced implementation of growth where growth is deemed adequate if and only if it is sufficient to lead to future proficiency.

Given their attachment to state achievement levels, growth-to-standard models tend to duplicate results provided by status models. In a study by Dunn (2007), 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. This is unsurprising, since students who are already proficient have much better chances of being on track to proficient than do non-proficient students. And, by extension, schools serving large proportions of non-proficient students minimally benefit when projected achievement is added to their already low achievement.

For a given school, the criterion-referenced growth-to-standard model yields a percentage of students on track to be proficient. This percentage confounds the present achievement level of the students at the school with the growth of the school's students. Specifically, schools with high percentages of students near or above the proficiency threshold will almost certainly possess higher percentages of students projected to be proficient than those schools with little or no proficient students. Assuming the growth-to-standard school percentages reflect the quality of the school (as their incorporation into accountability *vis-á-vis* AYP would), the results suggest that high achieving schools are almost always of higher quality than low achieving schools.

Due to the their close alignment with status—using growth to estimate future achievement—growth-to-standard models present a limited view of growth and serve, more generally, to impoverish the concept of growth as it relates to student achievement. To overcome this deficiency of growth-to-standard models, we contend that it is necessary to normatively embed these criterion-referenced growth methodologies. Given the evolution from norm- to criterion-referenced achievement, it seems logical that conditional achievement (i.e., growth) evolves similarly. The current effort to establish growth criteria absent growth norms runs counter to a half century of norm- and criterion-referenced achievement history. Following Angoff's (1974) dictum "scratch a criterion and you'll find a norm" we introduce *student growth percentiles* as a normative conceptualization of student growth.

Student Growth Percentiles

It is a common misconception that to measure student growth in education, the subject matter and grades over which growth is examined must be on the same scale—referred to as a vertical scale. Not only is a vertical scale not necessary, but its existence obscures fundamental concepts necessary to understand growth. Growth, fundamentally, requires change to be examined for a single construct, like math achievement, over time—*growth in what?* A single scale for the construct is necessary to measure the *magnitude* of growth, but not growth in general (Betebenner, 2008; Yen, 2007).

Consider the familiar situation from pediatrics where the interest is on measuring the height and weight of children over time. The scales on which height and weight are measured possess properties that educational assessment scales aspire toward but can never meet.

> An infant male toddler 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 an absolute magnitude of growth, but instead in a normative criterion locating that 4 inch increase alongside the height increases 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, 2008).

With this reality in the examination of change where scales of measurement are perfect, it seems absurd to think that in education, where scales are, at best, quasi-interval, one can/should examine growth differently.

Supposing scales did exist in education similar to height/weight scales that permitted the calculation of absolute measures of annual academic growth for students, the response parents receive to questions such as, "How much did my child progress?", would come as a number of scale score points—an answer likely to leave most parents bewildered wondering whether the number of points is good or bad. As in pediatrics, the search for a description regarding change in achievement over time (i.e., growth) is best served by considering a normative quantification of student growth—a student growth percentile.

The four panels of Figure 2 depict what a student growth percentile represents in a situation considering students having only two consecutive achievement test scores.

Upper left panel. Considering all pairs of scores for all students in the state yields a bivariate (two variable) distribution.

Upper right panel. Taking account of prior achievement (i.e., conditioning upon prior achievement) fixes a value of the 2005 scale score (in this case at 600) and is represented by the red slice taken out of the bivariate distribution.

Lower left panel. Conditioning upon prior achievement defines a *conditional distribution* which represents

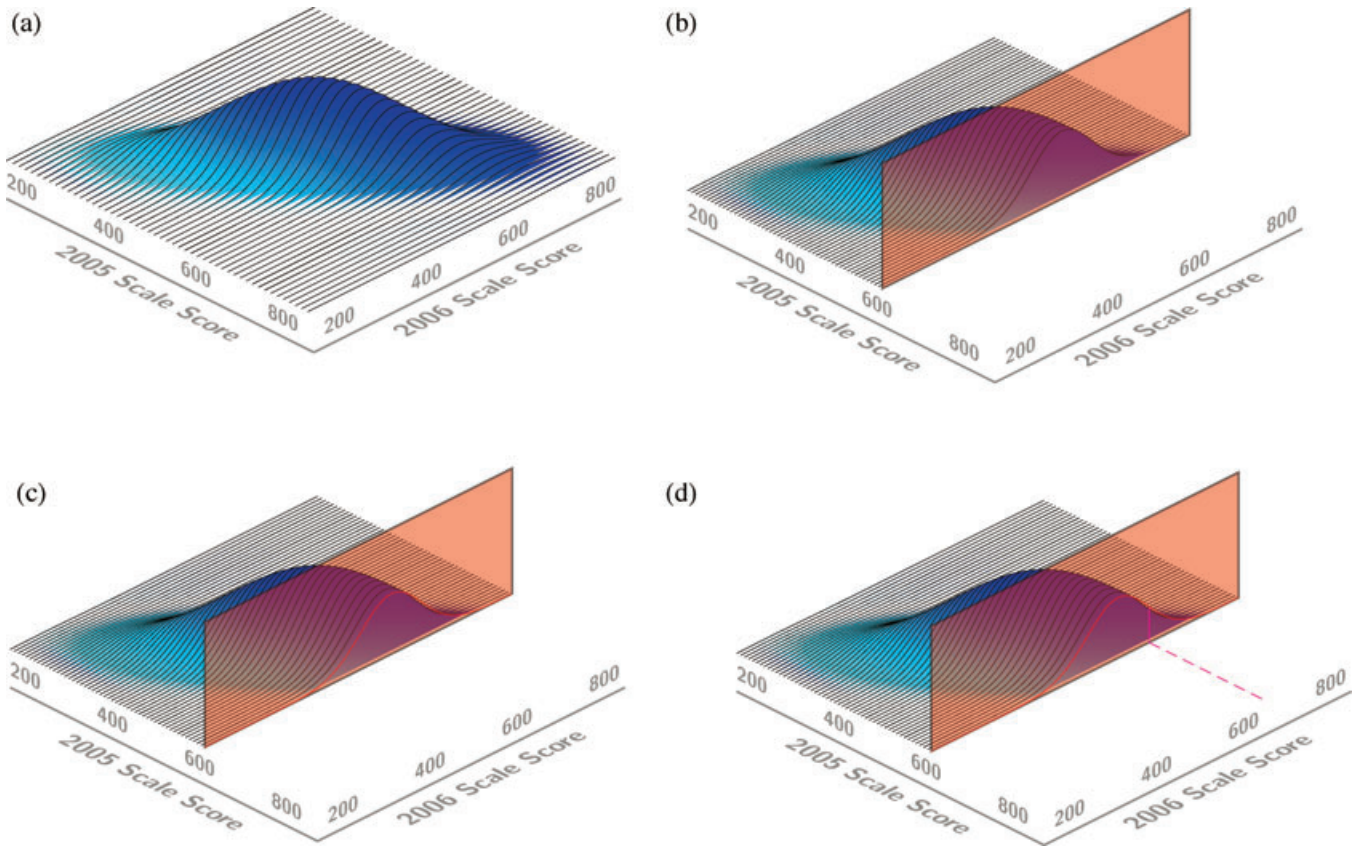

FIGURE 2. Figures depicting the distribution associated with 2005 and 2006 student scale scores together with the conditional distribution and associated growth percentile.

the distribution of outcomes on the 2006 test assuming a 2005 score of 600. This distribution is indicating as the solid red curve.

Lower right panel. The conditional distribution provides the context within which a student's 2006 achievement can be understood normatively. Students with achievement in the upper tail of the conditional distribution have demonstrated high rates of growth relative to their academic peers whereas those students with achievement in the lower tail of the distribution have demonstrated low rates of growth. Students with current achievement in the middle of the distribution could be described as demonstrating "average" or "typical" growth. In the figure provided the student scores approximately 650 on the 2006 test. Within the conditional distribution, the value of 650 lies at approximately the 70th percentile. Thus, the student's growth from 600 in 2005 to 650 in 2006 met or exceeded that of approximately 70% of students starting from the same place. This 50 point increase is above average. It is important to note that qualifying a student growth percentile as "adequate", "good", or "enough" is a standard setting procedure requiring stakeholders to examine a student's growth relative to external criteria such as performance standards/levels.

Figure 2 also illustrates the relationship between a vertical scale and student growth percentiles. Using the vertical scale implied by Figure 2, the student grew 50 points (from 600 to 650) between 2005 and 2006. This 50 points represents the magnitude of change. Quantifying the magnitude of change is scale-dependent. However, relative to other students, the achievement growth of the student has not changed—their growth percentile is invariant to scale transformations common in educational assessment. Student growth percentiles normatively situate achievement change bypassing questions associated with the magnitude of change toward the relative amount of change.

The percentile of a student's current score within their corresponding conditional distribution translates to a probability statement of a student obtaining a score at least/at most that level conditioning upon prior achievement. That is:[2]

$$\text{Student Growth Percentile} \equiv \Pr(\textit{Current Achievement} \mid \textit{Past Achievement}) \cdot 100. \quad (1)$$

Whereas unconditional percentiles normatively quantify achievement, conditional percentiles normatively quantify growth. Because past scores are used solely for conditioning purposes, one of the major advantages of using growth percentiles to measure change is that estimation does not require a vertical scale.

Calculation of conditional probability in Expression 1 does not require a common construct from past to present. It is, for example, possible to calculate the probability of current *math* achievement given prior *reading* achievement. This quantity, however, would be difficult to describe as "change" or "growth" along some developmental continuum. Some may argue that even with the same subject being used in Expression 1 that the belief in an unitary, underlying developmental continuum is tenuous at best. Avoiding the deeper philosophical questions of

the existence of such an underlying developmental continuum, we pursue the normative quantification of growth using student growth percentiles in contexts where the quantities from Expression 1 are useful in conversations about change in student achievement.

Student Growth Percentile Estimation

Calculation of a student's growth percentile is based upon the estimation of the conditional density associated with a student's score at time t using the student's prior scores at times $1, 2, \ldots, t-1$ as the conditioning variables. Given the conditional density for the student's score at time t, the student's growth percentile is defined as the percentile of the score within the time t conditional density. By examining a student's current achievement with regard to the conditional density, the student's growth percentile normatively situates the student's outcome at time t taking account of past student performance. The percentile result reflects the likelihood of such an outcome given the student's prior achievement. In the sense that the student growth percentile translates to the probability of such an outcome occurring (i.e., rarity), it is possible to compare the progress of individuals not beginning at the same starting point. However, occurrences being equally rare does not necessarily imply that they are equally "good." Qualifying student growth percentiles as "(in)adequate," "good," or as satisfying "a year's growth" is a standard-setting procedure requiring external criteria (e.g., growth relative to state performance standards) combined with the wisdom and judgments of stakeholders.

Estimation of the conditional density is performed using quantile regression (Koenker, 2005). Whereas linear regression methods model the conditional mean of a response variable Y, quantile regression is more generally concerned with the estimation of the family of conditional quantiles of Y. Quantile regression provides a more complete picture of both the conditional distribution associated with the response variable(s). The techniques are ideally suited for estimation of the family of conditional quantile functions (i.e., reference percentile curves). Using quantile regression, the conditional density associated with each student's prior scores is derived and used to situate the student's most recent score. Position of the student's most recent score within this density can then be used to qualify deficient/sufficient/excellent growth. Though many state assessments possess a vertical scale, such a scale is not necessary to produce student growth percentiles.

In analogous fashion to the least squares regression line representing the solution to a minimization problem involving squared deviations, quantile regression functions represent the solution to the optimization of a loss function (Koenker, 2005, p. 5). Formally, given a class of suitably smooth functions, $\mathcal{G}$, one wishes to solve

$$\arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(Y(t_i) - g(t_i)), \tag{2}$$

where t_i indexes time, Y are the time dependent measurements, and ρ_τ denotes the piecewise linear loss function defined by

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)) = \begin{cases} u \cdot \tau & u \geq 0 \\ u \cdot (\tau - 1) & u < 0. \end{cases}$$

The elegance of the quantile regression Expression 2 can be seen by considering the more familiar least squares estimators. For example, calculation of $\arg\min \sum_{i=1}^{n} (Y_i - \mu)^2$ over $\mu \in \mathbb{R}$ yields the sample mean. Similarly, if $\mu(x) = x\beta$ is the conditional mean represented as a linear combination of the components of x, calculation of $\arg\min \sum_{i=1}^{n} (Y_i - x_i\beta)^2$ over $\beta \in \mathbb{R}^p$ gives the familiar least squares regression line. Analogously, when the class of candidate functions $\mathcal{G}$ consists solely of constant functions, the estimation of Expression 2 gives the τth sample quantile associated with Y. By conditioning on a covariate x, the τth conditional quantile function, $Q_y(\tau \mid x)$, is given by

$$Q_y(\tau \mid x) = \arg\min_{\beta \in \mathbb{R}^p} \times \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta).$$

In particular, if $\tau = .5$, then the estimated conditional quantile line is the median regression line.[3]

Following Wei and He (2006), we parameterize the conditional quantile functions as a linear combination of B-spline cubic basis functions. B-splines are employed to accommodate non-linearity, heteroscedasticity, and skewness of the conditional densities associated with values of the independent variable(s). B-splines are attractive both theoretically and computationally in that they provide excellent data fit, seldom lead to estimation problems (Harrell, 2001, p. 20), and are simple to implement in available software.

Figure 3 gives a bivariate representation of linear and B-spline parameterization of decile growth curves. The assumption of linearity imposes conditions upon the heteroscedasticity of the conditional densities. Close examination of the linear deciles indicates slightly greater variability for higher grade 5 scale scores than for lower scores. By contrast, the B-spline-based decile functions better capture the greater variability at both ends of the scale score range together with a slight, non-linear trend to the data.

Calculation of student growth percentiles is performed using **R** (R Development Core Team, 2009), a language and environment for statistical computing, with **SGP** package (Betebenner, 2009). Other possible software (untested with regard to student growth percentiles) with quantile regression capability include SAS and Stata. Estimation of student growth percentiles is conducted using all available prior data, subject to certain suitability conditions. Given assessment scores for t occasions, ($t \geq 2$), the τ-th conditional quantile for Y_t based upon $Y_{t-1}, Y_{t-2}, \ldots, Y_1$ is given by

$$Q_{Y_t}(\tau \mid Y_{t-1}, \ldots, Y_1) = \sum_{j=1}^{t-1} \sum_{i=1}^{3} \phi_{ij}(Y_j)\beta_{ij}(\tau), \tag{3}$$

where $\phi_{i,j}$, $i = 1, 2, 3$ and $j = 1, \ldots, t-1$ denote the B-spline basis functions. Currently, bases consisting of 7 cubic polynomials are used to "smooth" irregularities found in the multivariate assessment data. A bivariate rendering of this is found is Figure 3 where linear and B-spline conditional deciles are presented. The cubic polynomial B-spline basis functions model the heteroscedasticity and non-linearity of the data to a greater extent than is possible using a linear parameterization.

The accuracy and precision of growth percentiles has not been formally investigated and is currently an active area of research. To frame the discussion of

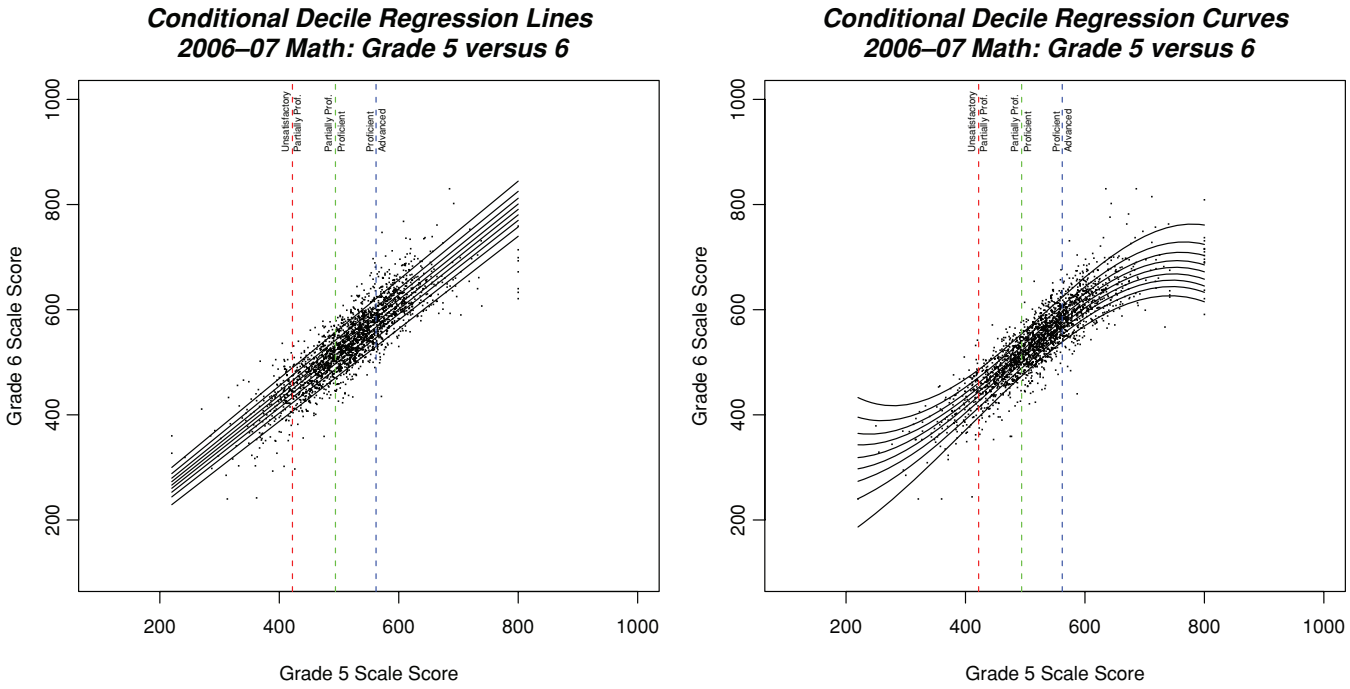

FIGURE 3. Linear and B-spline conditional deciles based upon bivariate math data, grades 5 and 6.

what is involved, it is necessary to decide among numerous potential chance process scenarios within which uncertainty is be quantified. The most important chance process at work is that associated with measurement of the individual and the error associated with that measurement: How precise and unbiased is the growth percentile estimate up to and including the measurement error of the individual? For example, given a student with a growth percentile of 90, how likely is it under parallel testing situations that the same student's growth percentile would be 50?[4]

How Much Growth Is Adequate?

Implicit in any normative description is an absence of criteria that, for example, could be used to qualify (in)adequacy. To address this shortcoming of student growth percentiles, it is necessary to establish adequacy criteria (i.e., standards) for student growth. Just as with criterion-referenced achievement, there are numerous ways in which this can be pursued. However, because achievement outcomes are of primary concern, the most natural way to create standards for growth is to quantify what level of growth is necessary to reach prescribed levels of achievement and use those results to inform a standard-setting procedure. This amounts to using a normative growth scale to enact a growth-to-standard approach.[5]

A significant impediment to the establishment of growth criteria (and the sensible use of growth analyses in general) is the imbroglio of terminology currently associated with discussions of student growth. Terms, found in Yen's earlier presented survey questions, like "a year's growth" are often not well defined. The following sections attempt to ground some of this terminology to enable better discussions about growth among stakeholders using growth percentiles as the basis for this discussion.

Defining Adequate Growth

To adequately address the notion of defining enough, adequate, or a year's growth, *aspirational* growth must be distinguished from *actual* growth:

Actual. What *is* a current year's growth?
Aspirational. What *should* a current year's growth be?

Answering the second question establishes a threshold distinguishing adequate from inadequate growth. To make such a distinction requires answering the first question which defines a norm: What is the range of growth currently observed? Aspirational growth for each student should be possible—this is Linn's existence proof applied at the individual level (Linn, 2003).

Student growth percentiles provide a means of answering the first question: What is a current year's growth? Answering the second question requires a qualification distinguishing adequate growth from inadequate growth. Using student growth percentiles, such a threshold can be established probabilistically or by growth to a standard (e.g., growth to a standard of proficiency within 3 years). The point of using the percentile scale for growth in this fashion is that the growth adequacy target has an immediate normative interpretation that can be used to set criterion-referenced aspirational goals that are reasonable.

Probabilistic adequacy. Perhaps the simplest (and least satisfying) way to define enough or adequate growth using growth percentiles is to stipulate a fixed growth percentile threshold that each student is required to meet or exceed. For example, a 50th percentile threshold (i.e., current typical growth) could be established to distinguish adequate from inadequate growth. As such, in any given year 50% of students would demonstrate adequate growth and 50% inadequate growth. An obvious feature/drawback of such a definition is that each year the same percentage of students will be categorized as having (in)adequate growth. This could potentially mask changes in growth rates over time, a phenomenon consistent with an education system becoming more or less effective.

Another disadvantage in setting target growth normatively is that it does not equalize chances for individuals to reach desired achievement outcomes (e.g., proficiency). That is, students well below proficient must demonstrate higher growth percentiles to reach proficient than students near the threshold. Similarly, students well above proficient need to demonstrate lower growth percentiles to maintain proficient than students near the threshold. This truism reflects fundamental criticism of defining adequacy in a normative fashion. If the ultimate goal is high achievement for all, then growth adequacy criteria will be student-specific—reflecting the current achievement level of the student.

Despite the disadvantages, there are advantages to establishing growth adequacy thresholds in a normative fashion. If the growth threshold (i.e., target growth) is defined uniformly for all students (e.g., 50th percentile growth), then there is probabilistic equivalence in terms of the difficulty of elevating each student to this growth target. This establishes an equitable goal for all students and, by aggregation, all schools. The drawback, as previously mentioned, is that this equitable growth goal, if achieved, does not lead all students to the equitable achievement goals that form the basis of accountability systems nationwide.

Despite being normative quantities, the methodology used to derive student growth percentiles can be leveraged to accommodate criterion-referenced aims. The most natural way to accomplish this is to tie growth adequacy to the achievement levels used in state accountability system. Doing so requires calculating what growth percentiles are necessary for students to reach the different achievement/performance level outcomes. These growth percentile goals can then be used to define growth adequacy thresholds that are both challenging and reasonable.

Growth-to-standard adequacy. To establish growth percentile targets (i.e., define what growth *should* be for each student) in terms of achievement levels, it is necessary to investigate what growth percentile is necessary to reach the desired performance level threshold based upon the student's achievement history. Intuitively, the lower one's scale score, the higher their growth percentile must be in order for them to reach the desired target. Equivalently, the lower one's current achievement the lower their chances of reaching the desired target. Specifically, if an individual must demonstrate 90th percentile growth to reach a desired achievement target (e.g., proficiency) in the coming year, then their chances of reaching such an outcome are .1 (i.e., 10%).

Establishing criterion-referenced growth thresholds requires consideration of multiple future growth/achievement scenarios. Instead of inferring that prior student growth is indicative of future student growth (e.g., linearly projecting student achievement into the future based upon past rates of change), predictions of future student achievement are contingent upon initial student status (where the student starts) and subsequent rates of growth (the rate at which the student grows). This avoids fatalistic statements such as, "Student *X* is projected to (not) be proficient in 3 years" and instead promotes discussions about the different rates of growth necessary to reach future achievement targets: "In order that Student *X* reach/maintain proficiency within 3 years, she will have to demonstrate *nth* percentile growth consecutively for the next 3 years." The change in phraseology is minor but significant. Stakeholder conversations turn from "where will (s)he be" to "what will it take?"

Figure 4 illustrates these parallel growth/achievement scenarios. Using the results of a growth percentile analysis based upon statewide longitudinal data, Figure 4 depicts five growth scenarios (10th, 35th, 50th, 65th, and 90th percentile growth), represented by the dark curves, in math for a student beginning in third grade at the break point between unsatisfactory and partially proficient. The figure depicts the four state achievement levels as shaded background regions across grades 3 to 10 together with the 2008 achievement percentiles across grades (inner most vertical axis) superimposed in white. The figure shows normative and criterion-referenced achievement simultaneously with normative and criterion-referenced growth.

Beginning at grade 3, a grade 4 achievement projection is made based upon the most recent growth percentile analyses derived using prior 3rd to 4th grade student progress. More specifically, using the coefficient matrices derived in the quantile regression of grade 4 on grade 3 (see Equation 3), predictions of what 10th, 35th, 50th, 65th, and 90th percentile growth yield are calculated. Next, using these five projected 4th grade scores combined with the starting 3rd grade score, 5th grade achievement projections are calculated using the most recent quantile regression of grade 5 on grades 3 and 4. The analysis extends and repeats for grades 6 to 10 yielding the *percentile growth trajectories* in Figure 4. The figures allow stakeholders to consider what *sustained* rates of growth that are observed presently yield for students starting at different points.

Figure 4 depicts math percentile growth trajectories for a student beginning at the threshold between achievement levels 1 and 2. Based upon the *achievement* percentiles depicted (white curves), approximately 7% of the population of 3rd graders are classified in achievement level 1. Following the white achievement percentile curve toward grade 10, the percentage of such students increases dramtically to near 35%. The black curves in the figure represent five different growth scenarios for the student based upon consecutive growth at a given growth percentile, denoted by the right axis. At the lower end, for example, consecutive 10th percentile growth leaves the student, unsurprisingly, mired in the lowest achievement category. Consecutive 35th, 50th, and 65th percentile growth also lead to a persistent unsatisfactory classification. This demonstrates how difficult (based upon current rates of progress) it is for students to move up in performance level in math statewide. With the second region from the top representing proficient, a student would need to demonstrate growth percentiles consecutively near 90 to reach this level of achievement by grade 10—showing how unlikely such an event currently is. In light of NCLB universal proficiency mandates, the growth necessary for the lower achieving students to reach proficiency, absent radical changes to growth rates of students statewide, is likely unattainable for a large percentage of these students.

Anchoring growth targets to achievement goals overcomes a previously mentioned shortcoming of a solely normative approach: The annually performed, normative growth percentile results are blind to any changes in the efficacy of the education system over time. If one of the ultimate goals of

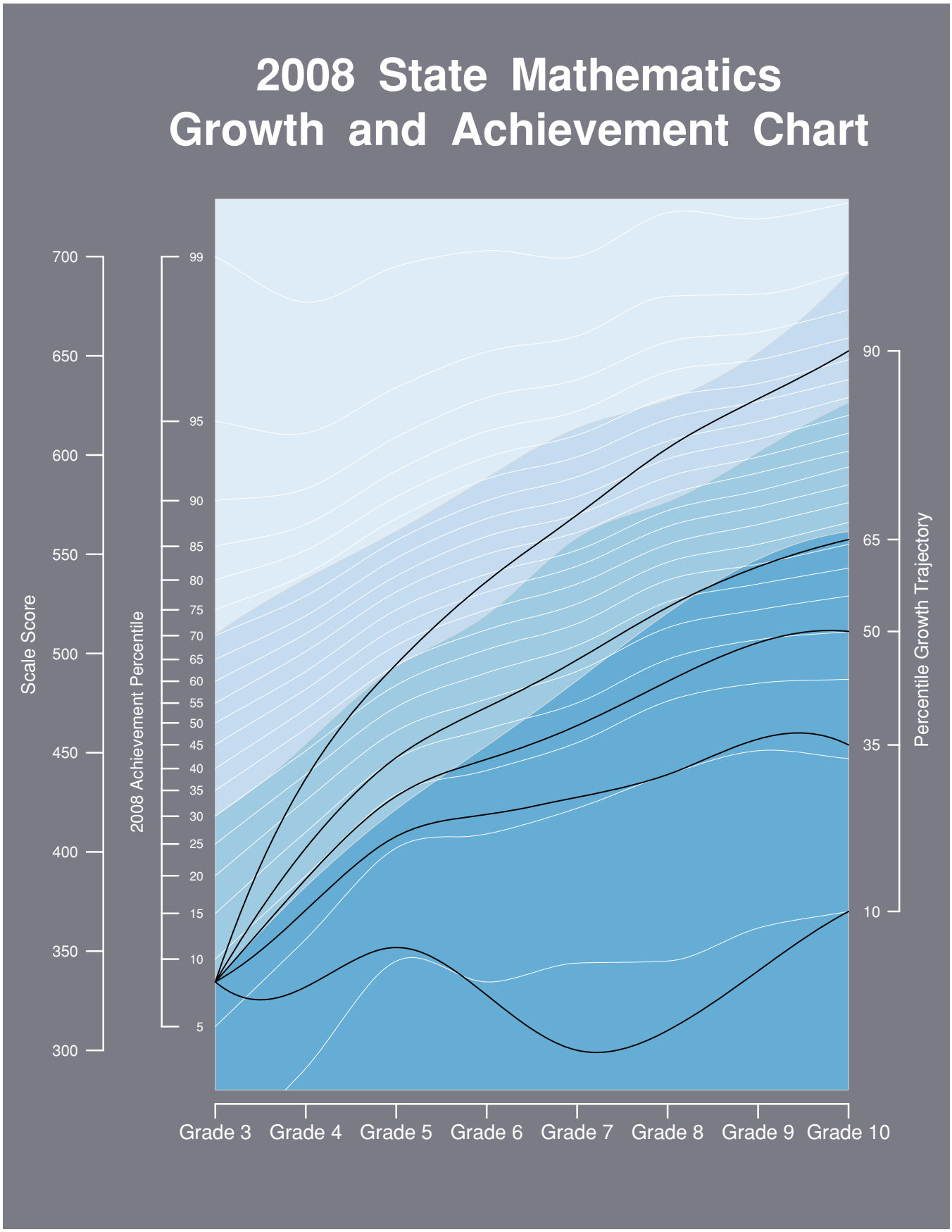

FIGURE 4. Math growth and achievement chart depicting norm- and criterion-referenced achievement (white lines and shaded background regions, respectively) across grades superimposed with five normative percentile growth trajectories (10th, 35th, 50th, 65th, and 90th) for a student beginning the third grade at the cutpoint between achievement levels 1 and 2.

education reform is to make today's 90th percentile growth tomorrow's 50th percentile growth, then to detect the system moving toward that goal, it is necessary to anchor the normative approach to some set of time-invariant standards. Using the achievement levels established for a state accountability system and the percentage of students reaching growth targets adequate to reach standards, given increasing educational efficacy, one would expect increasing percentages of students to reach these growth targets. It is important to note that, similar to comparing annual percent proficient results, such an approach relies heavily on the invariance of achievement levels and scale underlying them over time. Whether changes in educational efficacy are large enough to overcome error associated with equating and other sources of noise is not known.

One of the strengths of quantifying student growth normatively is that the growth percentile targets that are calculated of what it will take to reach a level of achievement quickly translate into the likelihood of such an event occurring. This dimension of student progress as it relates to accountability is absent from most growth-to-standard discussions. Today, achievement mandates are stipulated based upon the moral imperative of high standards for all children with little concern regarding the likelihood the students reaching these goals. Given current progress of students, it is unlikely that the sustained levels of growth necessary to reach these standards will occur soon.

Summary

The availability of longitudinal data from annual state assessment programs has created an unprecedented opportunity to examine the academic growth of students. Within the policy climate of the last decade modeling of student growth has focused primarily on assigning responsibility for growth. Even with recent flexibility toward growth modeling in federal accountability requirements, the combination of high-stakes accountability and universal proficiency mandates has impoverished the possibilities for applying growth analyses to annual state assessment data. As such, the rush to modify the criteria by which AYP for schools is determined has led states to consider student growth almost exclusively within the context of the accountability mandates of the original legislation.

Within this context, criterion-referenced growth (i.e., growth-to-standard) was used by a number of states in their applications for the GMPP. Growth-to-standard models make determinations about future achievement, designating whether students are on track to be proficient within a given time frame (usually 3 years). Such analyses are attractive from a policy-making perspective because they combine analyses of growth based upon scale scores with the NCLB universal proficiency mandates. However, such models are problematic in that they fail to adequately separate two essential qualities accountability systems wish to audit: achievement and effectiveness (Betebenner, 2006). Current NCLB performance mandates are achievement based—with a target of universal proficiency in achievement by 2014. Growth-to-standard results for schools are neither achievement measures nor student growth/effectiveness measures, but are an amalgam of the two which makes them difficult to interpret and use as a measure of school quality.

Box and Draper (1987, p. 74) famously asserted that, "All models are wrong, but some are useful." It is preferable to tailor a model to a set of desired uses rather than tailor uses to fit the strengths of a model. What, then, is the use stakeholders wish growth models to address? From preliminary surveys and work with states implementing growth models, stakeholders seem primarily interested in description. To this end, student growth percentiles are introduced as a means of satisfying this interest providing a quantification of student progress and a descriptive measure of *what is*. Criterion-referenced questions of *what should be* coincide with decisions about whether growth is "enough" or "adequate" to reach or maintain desired levels of achievement. The growth percentile metric serves to inform the standard setting procedure by communicating *what is possible*. In this author's opinion, only by considering simultaneously what is, what should be, and what is possible can accountability systems be equitable, just, and truly informed.

Notes

[1]The use of prior achievement as a consideration in qualifying current achievement is the most obvious but not the only choice of conditioning variable. Gender, race/ethnicity, socio-economic or special education status are potential candidates one might select to qualifying current status. Their use, however, is not justified in all cases. In *Educational Policy and the Just Society*, Strike (1982) distinguishes between morally relevant and irrelevant characteristics as they relate to describing achievement disparities. A morally relevant characteristic, for example, is prior achievement: where the child started. A morally irrelevant characteristic is the race/ethnicity of the child. Strike's distinction is *apropos* in considerations of what conditioning variables to consider in qualifying current student achievement.

[2]Technically, the expression denotes a student growth quantile since $\Pr(\textit{Current Achievement} \mid \textit{PastAchievement}) \cdot 100$ is not always an integer. To simplify, the result is rounded and termed a percentile.

[3]For a detailed treatment of the procedures involved in solving the optimization problem associated with Expression 2, see Koenker (2005), particularly Chapter 6.

[4]Other chance scenarios that can be considered involve the linking of tests across years as well as issues of uncertainty arising from regression analyses. The impact of linking error on growth percentiles is unknown but worth considering. Uncertainty arising from regression analyses depends upon whether the user considers the norming group to be a sample or a population. In situations involving growth percentile analyses with state assessment data, the construction of a superpopulation containing the state's population seems contrived at best (Berk, 2003). The regression employed in the calculation of growth percentiles is used to *descriptively* relate independent and dependent variables within a population of interest and is not used for *inferential* purposes with regard to some superpopulation (Berk, 2003).

[5]This approach was used by Colorado in their recently (January, 2009) approved application for the GMPP to use growth within state AYP determinations.

References

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, *92*, 2–5.

Auty, W., & Group, A. W. (2008). *Implementer's guide to growth models* (Tech. Rep.). Washington, DC: Council of Chief State School Officers (CCSSO). Retrieved December 10th, 2008 from http://www.ccsso.org/content/pdfs/IGG\%20Final\%20AP.pdf.

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

Berk, R. A. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Betebenner, D. W. (2006, January). *Growth as a description of process*. Paper presented at the Festschrift dedicated to the

life and work of Robert L. Linn, sponsored by the National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

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: Taylor & Francis.

Betebenner, D. W. (2009). ***SGP**: Student growth percentile and percentile growth projection/trajectory functions* (R package version 0.0-4). Vienna, Austria: R Development Core Team.

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

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models* (Tech. Rep.). Princeton, NJ: Educational Testing Service.

Carlson, D. (2001). *Focusing state educational accountability systems: Four methods for judging school quality and progress*. Retrieved Sept 18th, 2005 from http://www.nciea.org.

Dunn, J. (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 SCASS, Nashua, NH.

Edley, C. (2006, April 10). *Educational "opportunity" is the highest civil rights priority. so what should researchers and lawyers do about it?* [Video of speech with slides posted on the World Wide Web]. Retrieved June 22, 2006 from the World Wide Web: http://www.softconference.com/MEDIA/WMP/260407/\#43.010.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer.

Harris, D. N. (2007). *The policy uses and "policy validity" of value-added and other teacher quality measures* (Tech. Rep.). Princeton, NJ: Educational Testing Service.

Hill, R. K. (2002, April). *Examining the reliability of accountability systems*. (Paper presented at the 2002 meeting of the American Educational Research Association, New Orleans. Retrived August 10th, 2006 from http://www.nciea.org/publications/NCME_RHCD03.pdf.

Hill, R. K., & DePascale, C. (2002). *Determining the reliability of school scores* (Tech. Rep.). Washington, DC: Council of Chief State School Officers (CCSSO). Retrived July 10th, 2006 from http://www.ccsso.org/content/pdfs/DeterminingReliability.pdf.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations* (Tech. Rep.). Los Angeles, CA: Center for the Study of Evaluation, CRESST.

Linn, R. L. (2008). Educational accountability systems. In *The future of test-based educational accountability* (pp. 3–24). New York: Taylor & Francis.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

R Development Core Team (2009). ***R**: A language and environment for statistical computing*. Vienna, Austria: Author (3-900051-07-0).

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), 121–129.

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

Spellings, M. (2005). *Secretary Spellings announces growth model pilot* [Press Release]. Washington, DC: U.S. Department of Education. Retrieved August 7, 2006 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html.

Strike, K. (1982). *Educational policy and the just society*. Chicago: University of Illinois Press.

Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics*, *34*(5), 2069–2097.

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: Springer.