Norm- and Criterion-Referenced Student Growth[†]

Damian W. Betebenner
National Center for the Improvement of Educational Assessment
Dover, NH
DBetebenner@nciea.org

March 20, 2008

Abstract

Annual student achievement data derived from state assessment programs have led to wide spread enthusiasm for statistical models suitable for longitudinal analysis. In response, the United States Department of Education recently solicited growth model proposals from states as a means of satisfying NCLB adequate yearly progress requirements. Given the current policy environment's rigid adherence to NCLB's universal proficiency mandate, the preponderance of models thus far proposed maintain compliance by estimating future (i.e., projected) student achievement. Referred to as the "growth-to-standard" approach, these criterion referenced growth models designate whether a student is "on track to being proficient" and use this designation as evidence of school quality. This paper begins by situating current growth-to-standard approaches within a larger domain of statistical models including those based solely upon achievement as well as more traditional growth models. Within this context, we demonstrate that current growth-to-standard approaches present an impoverished view of student progress because they lack a normative foundation. To remedy this, student growth percentiles are introduced as a normative description of growth capable of accommodating, informing, and extending criterion referenced aims like those embedded within NCLB.

Background

Accountability systems constructed according to federal adequate yearly progress (AYP) requirements currently rely upon annual measurement 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). Status measures, though appropriate for making judgments about the achievement level of students at a school for a given year, 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 NCLB to label its accountability provisions as unfair and misguided and to demand the use of growth analyses as a better means of auditing the quality of schools.

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 models into state 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).

In guidance to states applying for the GMPP, the United States Department of Education stated explicitely 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. In response, of the models thus far approved as part of the GMPP, a majority maintain compliance by examining "growth" in terms of future (i.e., projected) student achievement. Referred to as the "growth-to-standard" approach, these criterion referenced growth

[†]An electronic version of this paper is available at http://www.nciea.org.

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.[1] Operationalizing growth as projected achievement represents a departure from more familiar student growth models, including the widely discussed value-added models.

The primary thrust of growth analyses over the last decade has been to determine, using sophisticated statistical techniques, the amount of student progress/growth that can be justifiably attributed to the school or teacher (Braun, 2005; Rubin, Stuart, & Zanutto, 2004; Ballou, Sanders, & Wright, 2004; Raudenbush, 2004). Such analyses, often called *value-added* analyses, attempt to estimate the teacher/school contribution to student achievement. This contribution, called the *school* 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 this utility, such analyses fail to address one of the fundamental questions stakeholders have regarding student growth: Namely, how much growth did a student make? In line with this question, the purpose of this paper is to present a third way between value-added models and the growth-to-standard approach. This paper addresses the fundamental task of quantifying how much growth a student makes. Borrowing concepts from pediatrics used to describe infant/child weight and height progressions, this paper introduces *student growth percentiles* (Betebenner, 2008). These individual reference percentiles side step 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. Student growth percentiles can be used to understand growth both normatively as well as in a criterion-referenced fashion *vis-à-vis* current growth-to-standard approaches. With regard to the latter, growth percentile methodology can be used to answer the question of how much growth is necessary for a student to attain/maintain proficiency within an allotted timeframe.

Status and Growth

The impact of NCLB upon research connecting large scale assessment outcomes and 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; Linn et al., 2002; Hill, 2002; Hill & DePascale, 2002; Carlson, 2001). The rigid semantical distinction between status and growth models obscures common foundation: namely, to understand and ultimately qualify student achievement.[2] What considerations, if any, are necessary to understand a student's 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 status (i.e., achievement level) of the student. Hence, 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 simple qualification of achievement for each student based upon the state's performance standards. As the basis for an accountability system with rigorous achievement standards, such models are extremely demanding, requiring, without condition, an acceptable level of achievement from all students.

A natural extension to the qualifications 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? Is their current level of achievement exemplary given where they were? Conditional status models, or growth models, evaluate student progress based upon a longitudinal record of student achievement.[3] Figure 1 depicts the distinction of growth versus status as a difference between whether

[1] 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).

[2] The term qualify is used in the sense of rendering a judgment about the (in)adequacy of an observed performance.

[3] 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

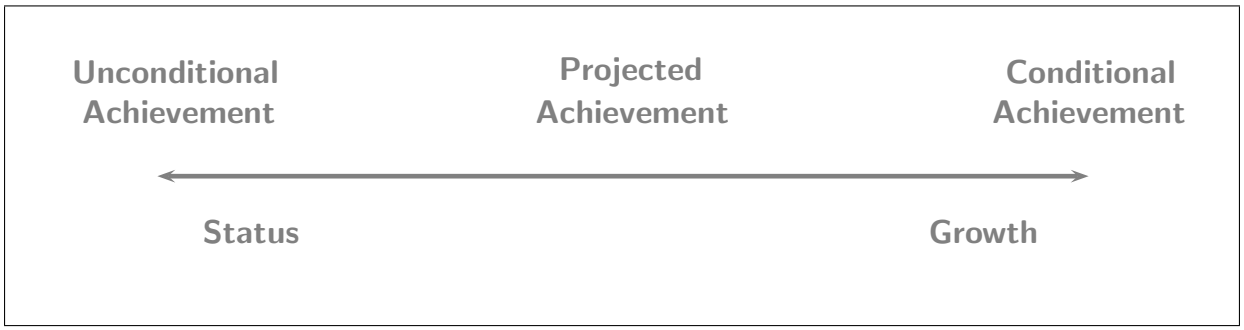

Figure 1: Unconditional, projected, and conditional achievement and their relationship to status and growth

or not achievement is examined unconditionally or conditionally.

Situated between growth and status in Figure 1 are the projected achievement (i.e., growth-to-standard) models, the current model of choice for accommodating NCLB achievement mandates. Using various statistical procedures, the models predict the future achievement of the student (usually up to three 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 status to make a determination about whether the student's growth is adequate.[4] 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 close attachment to state performance levels, growth-to-standard models tend to simply repackage many of the characteristics already present in 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 projected to be 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. Consequently, due to the their close alignment with status—using growth to estimate future achievement—growth-to-standard models represent a distorted 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 methodologies, a task consistent with an aphorism attributed to Bob Linn: "scratch a criterion and you'll find a norm". Given the evolution from norm- to criterion-referenced achievement, it seems logical that conditional achievement (i.e., growth) evolve similarly. The current effort to establish growth criteria absent growth norms runs counter to a half century of norm- and criterion-referenced achievement history. To remedy this situation, we introduce *student growth percentiles* as a normative context in which to ground student growth. We demonstrate that once a normative basis for growth is established, discussions regarding what constitutes "adequate growth", "a year's growth" or "enough growth" reduce to standard setting procedures *vis-à-vis* achievement standard setting procedures.

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

consider in qualifying current student achievement.

[4]Somewhat paradoxically, all growth-to-standard models the author has reviewed qualify a student's growth as (in)adequate without actually quantifying how much the student did grow.

assessment scales aspire towards 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.

With this reality in the examination of change where scales of measurement are perfect, it is 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.

A student's growth percentile describes how (ab)normal a student's growth is by examining their current achievement relative to their *academic peers*—those students with identical prior achievement. That is, a student growth percentile examines the current achievement of a student relative to other students who have, in the past, "walked the same achievement path". Heuristically, if the state assessment data set were extremely large (in fact, infinite) in size, one could examine the data set and select out those students with the exact same prior scores and compare how the selected student's current year score compares to the current year scores of those students with the same prior year's scores—their academic peers. If the student's current year score exceeded the scores of most of their academic peers, in a normative sense they have done well. If the student's current year score was less than the scores of their academic peers, in a normative sense they have not done well.

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 the 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 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 percent 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.

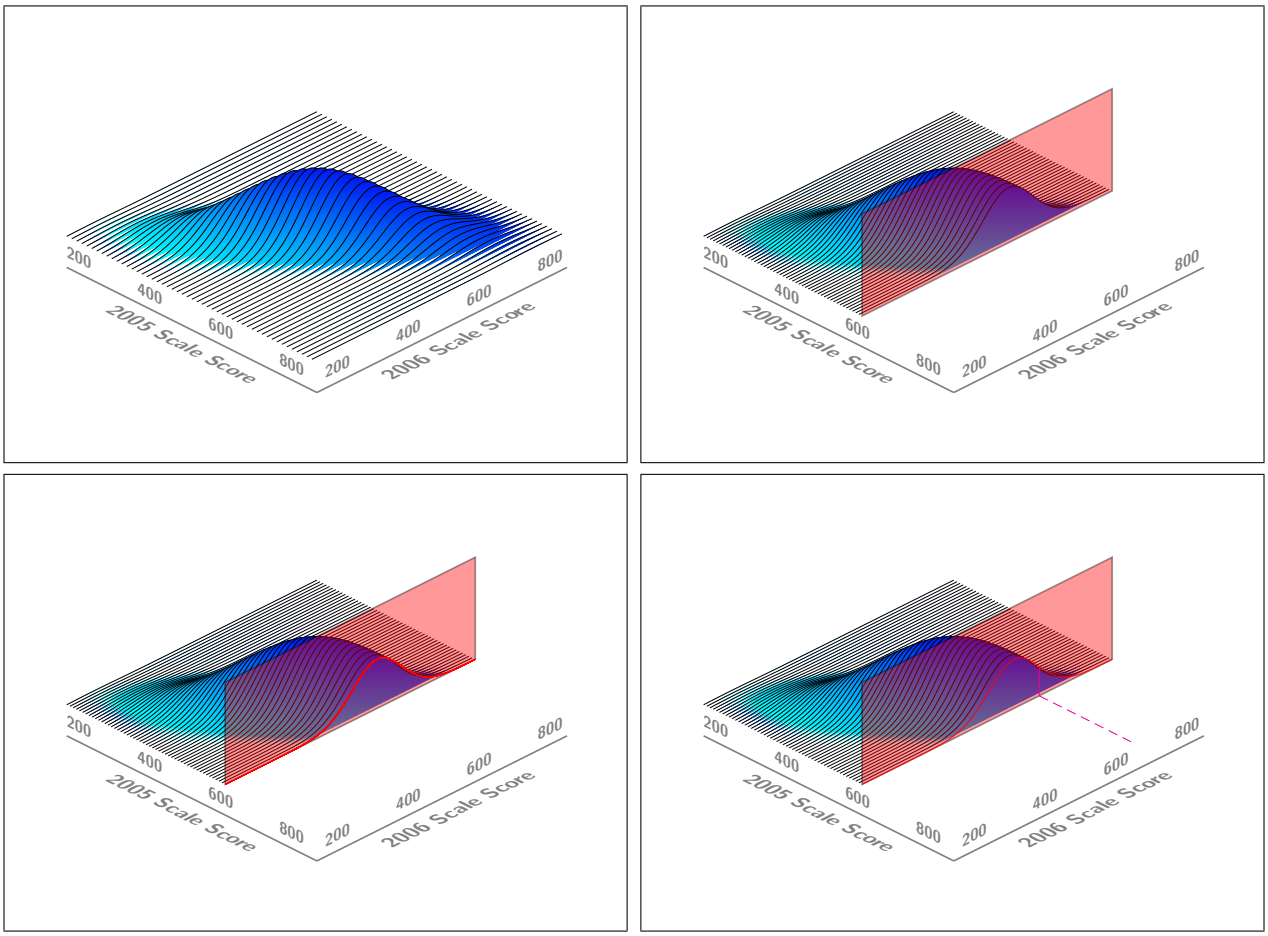

Figure 2: Figures depicting the distribution associated with 2005 and 2006 student scale scores together with the conditional distribution and associated growth percentile

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, and directing attention toward relative standing which, we believe, is likely to interest stakeholders most.

The percentile of a student's current score within their corresponding conditional distribution translates to a probability statement of a student obtaining that score taking account of prior achievement. That is:[5]

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

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.

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

[5] Technically, the expression denotes a student growth quantile since $\Pr(\textit{Current Achievement}|\textit{Past Achievement}) \cdot 100$ is not always an integer between 1 and 100. To simplify, the result is rounded down and termed a percentile.

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

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

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 1 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 1 gives the τth sample quantile associated with Y. By conditioning on a covariate x, the τth conditional quantile function, $Q_y(\tau|x)$, is given by

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

In particular, if $\tau = 0.5$, then the estimated conditional quantile line is the median regression line.[6]

Following Wei & 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-splines 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.

[6] For a detailed treatment of the procedures involved in solving the optimization problem associated with Expression 1, see Koenker (2005), particularly Chapter 6.

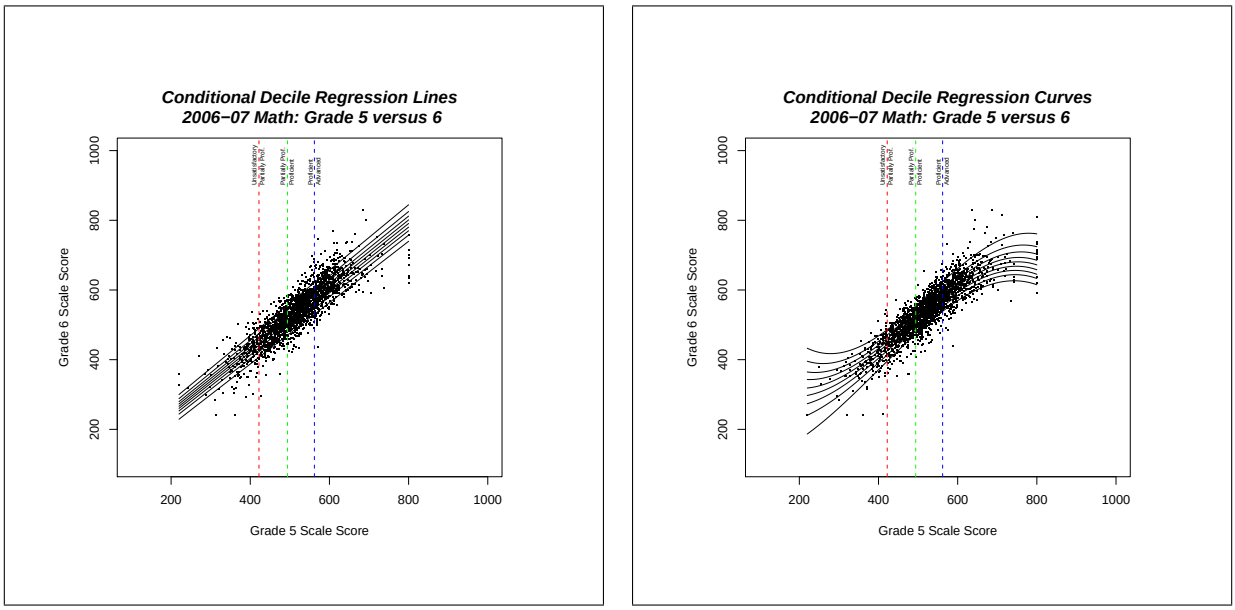

Figure 3: Linear and B-spline conditional deciles based upon bivariate math data, grades 5 and 6

Currently, calculation of student growth percentiles is performed using R, a language/environment for statistical computing, with Koenker's `quantreg` package (R Development Core Team, 2006). 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

$$Q_{Y_t}(\tau|Y_{t-1}, \ldots, Y_1) = \sum_{j=1}^{t-1} \sum_{i=1}^{3} \phi_{ij}(Y_j)\beta_{ij}(\tau), \tag{2}$$

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.

Discussion of Model Properties

Student growth percentiles possess a number of attractive properties from both a theoretical as well as a practical perspective. Foremost among practical considerations is that the percentile descriptions are familiar and easily communicated to teachers and other non-technical stakeholders. Furthermore, implicit within the percentile quantification of student growth is a statement of probability. Questions of "how much growth is enough?" or "how much is a year's growth?" ask stakeholders to establish growth percentile thresholds deemed adequate. These thresholds establish growth standards that translate to probability statements. In this manner, percentile based growth forms a basis for discussion of rigorous yet attainable growth standards for all children supplying a normative context for Linn's (2003) existence proof with regard to student level growth.

In addition to practical utility, student growth percentiles possess a number of technical attributes well suited for use with assessment scores. The more important theoretical properties of growth percentiles include:

Robustness to outliers Estimation of student growth percentiles are more robust to outliers than is traditionally the case with conditional mean estimation. Analogous to the property of the median being less influenced by outliers than is the median, conditional quantiles are robust to extreme observations. This

is due to the fact that influence of a point on the τ-th conditional quantile function is not proportional (as is the case with the mean) to the distance of the point from the quantile function but only to its position above or below the function (Koenker, 2005, p. 44).

Uncorrelated with prior achievement Analogous to least squares derived residuals being uncorrelated with independent variables, student growth percentiles are not correlated with prior achievement. This property runs counter to current multilevel approaches to measuring growth with testing occassion nested within students (Singer & Willett, 2003). These models, requiring a vertical scale, fit lines with distinct slopes and intercepts to each student. The slopes of these lines represent an average rate of increase, usually measured in scale score points per year, for the student. Whereas a steeper slope represents more learning, it is important to understand that using a normative quantification of growth, one cannot necessarily infer that a low achieving student with a growth percentile of 60 "learned as much" as a high achieving student with the same growth percentile. Growth percentiles bypass questions associated with magnitude of learning and focus on normatively quantifying changes in achievement.

Equivariance to monotone transformation of scale An important attribute of the quantile regression methodology used to calculate student growth percentiles is their invariance to monotone transformations of scale. This property, denoted by Koenker (2005) as *equivariance to monotone transformations* is particularly helpful in educational assessment where a variety of scales are present for analysis, most of which are related by some monotone transformation. For example, it is a common misconception that one needs a vertical scale in order to calculate growth. Because vertical and non-vertical scales are related via a monotone transformation, the student growth percentiles do not change given such alterations in the underlying scale. This result obviates much of the discussion concerning the need for a vertical scale in measuring growth.[7]

Formally, given a monotone transformation h of a random variable Y,

$$Q_{h(Y)|X}(\tau|X) = h(Q_{Y|X}(\tau|X)).$$

This result follows from the fact that $\Pr(T < t|X) = \Pr(h(T) < h(t)|X)$ for monotone h. It is important to note that equivariance to monotone transformation does not, in general, hold with regard to least squares estimation of the conditional mean. That is, except for affine transformations h, $E(h(Y)|X) \neq h(E(Y|X))$. Thus, analyses built upon mean based regression methods are, to an extent, scale dependent.

Student growth percentiles derived using quantile regression procedures possess a number of attractive properties that make them ideal candidates as normative descriptors of student growth. An obvious criticism of growth percentiles, as well as any other normatively derived quantity, is that it is purely descriptive and inappropriate for NCLB like determinations requiring adequacy judgments. The next section confronts this criticism and shows that growth percentiles and the methodology underlying them are ideally suited for tasks involving criterion reference growth.

How much Growth is Adequate?

Having described how to normatively quantify growth in terms of student growth percentiles, we now turn our attention and demonstrate how student growth percentiles can form the basis of growth-to-standard discussions regarding what is adequate' growth. This is the process of going from a norm-referenced to a criterion-referenced standard of growth and is, in many ways, analogous to current normative and criterion referenced understandings of achievement. We begin by unpacking the imbroglio of terminology currently associated with discussion of student growth.

In a discussion of growth, NCLB, and vertical scaling, Yen (2007), provides a list of questions regarding growth taken from a survey of parents, teachers, and administrators:

Parent Questions:

- Did my child make a year's worth of progress in a year?

[7] As already noted with regard to pediatrics, the existence of nice "vertical" scales for measuring height and weight still leads to observed changes being normed.

- 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 concludes, all these questions rest upon a desire to understand whether observed student progress is "reasonable or appropriate" (Yen, 2007, p. 281). Moreover, the questions admit two paths to their resolution: the absolute and the normative. As discussed previously, student growth percentiles provide a normative and, we more informative way to address these questions than an absolute metric on which to interpret growth.

Methodology for 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—again, Linn's existence proof applied at the individual level (Linn, 2003).

Student growth percentiles provide an elegant 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. For example, the current growth-to-standard criterion utilized by most states in the GMPP defines adequate growth as growth leading toward proficiency. Using the conditional densities like those depicted in Figure 2, it is straightforward to calculate the growth percentile necessary for each student to reach that level of achievement. This threshold could then be used to distinguish adequate from inadequate growth. A benefit of using the percentile scale for growth is that the threshold has a normative context that can be used to set criterion referenced aspirational goals that are reasonable.

Norm-referenced adequacy

Perhaps the simplest 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 used to distinguish adequate from inadequate growth. Given present circumstances, 50% of students would be expected to demonstrate adequate and 50% inadequate growth. Going forward, relative to growth demonstrated by the baseline cohort, it is possible that more than 50% of students could demonstrate growth exceeding the baseline established 50th percentile. Such

a phenomenon would be consistent with the goal for greater effectiveness within the educational system—for today's exemplary growth to become tomorrow's average growth.

There are advantages to establishing growth adequacy thresholds in a normative fashion. If the growth threshold (i.e., target growth) is defined uniformly for each student (e.g., establishing target growth at the 50th growth percentile from baseline data), then there is probabilistic equivalence in terms of the difficulty of elevating each student to this growth target. If percentages of students at a school achieving target growth is reported, then the goal for each school is to get students to grow at a rate exceeding the baseline 50th percentile growth. This establishes an equitable goal for all students and schools. If students were randomly distributed to schools, then in the baseline year, 50% of the students in any given school would be expected to demonstrate target growth or better. As time passes, one hopes to observe schools having greater and greater percentages of their students in this baseline category.

A disadvantage to setting target growth normatively is that it doesn't equalize chances for individuals to reach proficiency (or other achievement outcomes associated with state defined performance levels). This represents a fundamental criticism of defining adequacy in a normative fashion. However, even though student growth percentiles are normative measures of student growth they are in no way constrained by normative adequacy criteria. Growth percentiles derived with quantile regression can be used to establish growth standards in terms of performance levels. To do so requires investigating 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.

Criterion-referenced adequacy

To establish growth percentile targets (i.e., define what growth *should* be for each student) in terms of performance 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 0.1 (i.e., 10 percent).

Establishing criterion referenced growth thresholds requires consideration of multiple future 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). Instead of fatalistic statements such as, "Student X is projected to be (not) proficient in three years", discussions that hypothesize different rates of growth and their consequences with regard to future achievement are considered: "Given that Student X starts at this point and grows over the coming three years at rate G, we anticipate Student X will (not) be proficient." The change is phraseology is minor but significant. Stakeholder conversations turn from "where will (s)he be" to "what will it take?"

Parallel growth scenarios are more easily understood with the assistance pictures. Figures 4 to 12 (Pages 15 to 23) depict three growth scenarios each in math, reading, and writing for students beginning in third grade at each of three performance level cutpoints (i.e., between unsatisfactory/partially proficient, partially proficient/proficient, and proficient/advanced). The figures depict the four state performance levels across grades 3 to 10 in color together with the 2007 achievement percentiles superimposed in white. Beginning at grade 3 at the given cutpoint, a grade 4 achievement projection is made based upon the the growth percentile derived using prior 3rd to 4th grade student progress. Next, using this projected 4th grade score combined with the 3rd grade score, a 5th grade achievement projection is made using prior student progress from 3rd and 4th to 5th. The process repeats to plot out different "growth percentile trajectories". The figures allow stakeholders to consider what 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth (sustained year-over-year) yields for students with three hypothetical starting points in the 3rd grade. Like all forecasting, these projections are not exact, especially as the timeframe extends. However, the charts do allow for a "bird's eye view" that can aid stakeholders in growth standard setting.

Consider Figure 4, math growth trajectories for a student beginning at the unsatisfactory/partially proficient threshold. Based upon the achievement percentiles, approximately 7 percent of the population of 3rd graders rate as unsatisfactory. Moving toward grade 10, the percentage of unsatisfactory students increases

dramtically to near 35 percent. The black lines in the figure represent six 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 25th percentile growth leaves the student, unsurprisingly, mired in the unsatisfactory category. Consecutive 40th, 50th and 60th percentile growth also leave the child in the unsatisfactory category. This demonstrates how difficult (based upon current rates of progress) it is for students to move up in performance level in math. With the green region representing proficient, a student would need to demonstrate growth percentiles in excess of 75 to reach proficiency showing how unlikely such a event currently is. Considering NCLB universal proficiency mandates, the growth necessary for non-proficient students to reach proficiency is likely unattainable for a large percentage of non-proficient students given current levels of growth. Of course, the reality of the present need not define a blueprint for the future. However, without a radical restructuring of math education for those non-proficient students, it seems highly unlikely that the achievement targets of NCLB will be realized in this education system.

If the goal of an accountability system is universal proficiency, then the growth percentile targets can be set accordingly. One of the strengths of quantifying student growth normatively is that growth percentile targets quickly translate into the likelihood of such an event occurring. This dimension of student improvement 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. Given current progress of students, it is unlikely that the sustained levels of growth necessary to reach these standards will ever occur. A fundamental dictum of moral philosophy ascribed to Kant is that "ought implies can": If someone *ought* to do something, they can do it, in the sense that they have the possibility/capacity to do it (Betebenner & Howe, 2007). Growth percentiles bring Kant's dictum to the fore when considering criterion referenced growth standards.

Discussion

Recent flexibility in federal accountability requirements has given states the opportunity to augment their current status based accountability systems with growth analysis techniques—taking account of student progress over time. In November 2005, Secretary of Education Spellings announced the GMPP, allowing states to submit proposals to supplement accountability systems with some form of student growth. Proposal requirements mandated that states maintain the original 2014 universal achievement mandates. The rush to modify the criteria by which adequate yearly progress (AYP) for schools is determined has led states to submit models that fail to address significant shortcomings of the accountability mandates of the original legislation.

Toward these ends, the growth-to-standard approach was used by a number of states in their applications for the GMPP. Growth-to-standard models make projections about future achievement and use these projections to designate whether students are on track to be proficient within a given time frame (usually three years). Such analyses are attractive from a policy making perspective because they combine analyses of growth based upon scale scores with the universal proficiency mandates on which NCLB accountability systems rest. However, such models are problematic in that they fail to adequately distinguish between the 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 instead a mixture of the two which makes them difficult to interpret and use as a measure of school quality.

Though the growth-to-standard conceptualization begins with individual student growth model, the aggregate results derived from such models for schools (or teachers or districts) fail to provide a defensible measure of student growth or effectiveness. For a given school, the models yield a percentage of students projected to be proficient. However, this percentage confounds the present achievement level of the students at the school with the 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 quantify effectiveness, the results would suggest that high achieving schools are almost always more effective than low achieving schools. Not surprisingly, preliminary results from states implementing such models as a means of satisfying AYP have found little overall change in the number of schools that benefit from this type of growth analysis (Dunn, 2007).

If policy makers wish to build accountability systems that incorporate individual student growth, then it

is imperative to understand the distinction between individual growth and system level effectiveness. It is not uncommon to hear phrases such as "student growth" or "school growth" used without serious consideration of what, for example, "school growth" means. Does "school growth" refer to the aggregate growth rate of students in a given school or does it refer to changes (i.e., aggregate level change) in school level effectiveness? The two represent fundamentally different, but related, concepts each of which provide information about the quality of the education system (Betebenner, 2008; Carlson, 2001). Where, we would argue, policies like NCLB are concerned with increasing education effectiveness, it has not been well articulated how this "trickles down" to expectations for individual students.[8]

This paper asserts that an unavoidable step toward the incorporation of student growth into criterion referenced accountability systems is a normative understanding of student growth. To this end, student growth percentiles estimated using quantile regression techniques are introduced as a means of fulfilling this requirement. Student growth percentiles provide a descriptive measure of *what is*, that is, a quantification of how much a student grew. Questions of *what should be* coincide with decisions about whether growth is "enough" or "adequate". These qualifications rest outside of statistics and require reasoned judgment on the part of stakeholders to set such standards. The student growth percentile metric serves to inform the standard setting procedure by immediately relaying *what is possible*. Only by considering, what is, what should be, and what is possible simultaneously, can growth standard setting and accountability systems built upon such standards be equitable, just, and truly informed.

References

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.

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 L. A. Shepard & K. Ryan (Eds.), *Festschrift in honor of the life and work of Robert L. Linn.* Thousand Oaks, CA: Sage Publications.

Betebenner, D. W., & Howe, K. R. (2007). *On NCLB, equal educational opportunity and closing achievement gaps.* (Unpublished. To appear in Educational Researcher)

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

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

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, September 25, 2007, Nashua, NH)

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

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

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.

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

[8] Though NCLB status based accountability requirements focus on achievement, they mandate increasing achievement over time. Unless there are changes to the underlying student population being educated, increasing achievement rates are likely an indicator of increasing effectiveness.

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

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. (2006). *R: A language and environment for statistical computing.* Vienna, Austria. (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.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis.* New York: Oxford University Press.

Spellings, M. (2005, Nov). *Secretary Spellings announces growth model pilot* [Press Release]. 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 & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York: Springer.

Growth Projection Figures

The growth projection figures that follow present 9 different growth scenarios (3 in reading and 3 in math) for students with 3rd grade state assessment scores at the performance level cutpoints. The growth scenarios depict consecutive (i.e., year-over-year) growth quantified in terms of student growth percentiles. That is, the figures present what, for example, 50th percentile growth leads to versus 60th percentile growth. The figures are intended to aid stakeholders in better understanding the range of student growth and what different growth rates lead to in terms of student achievement relative to state designated performance performance levels.

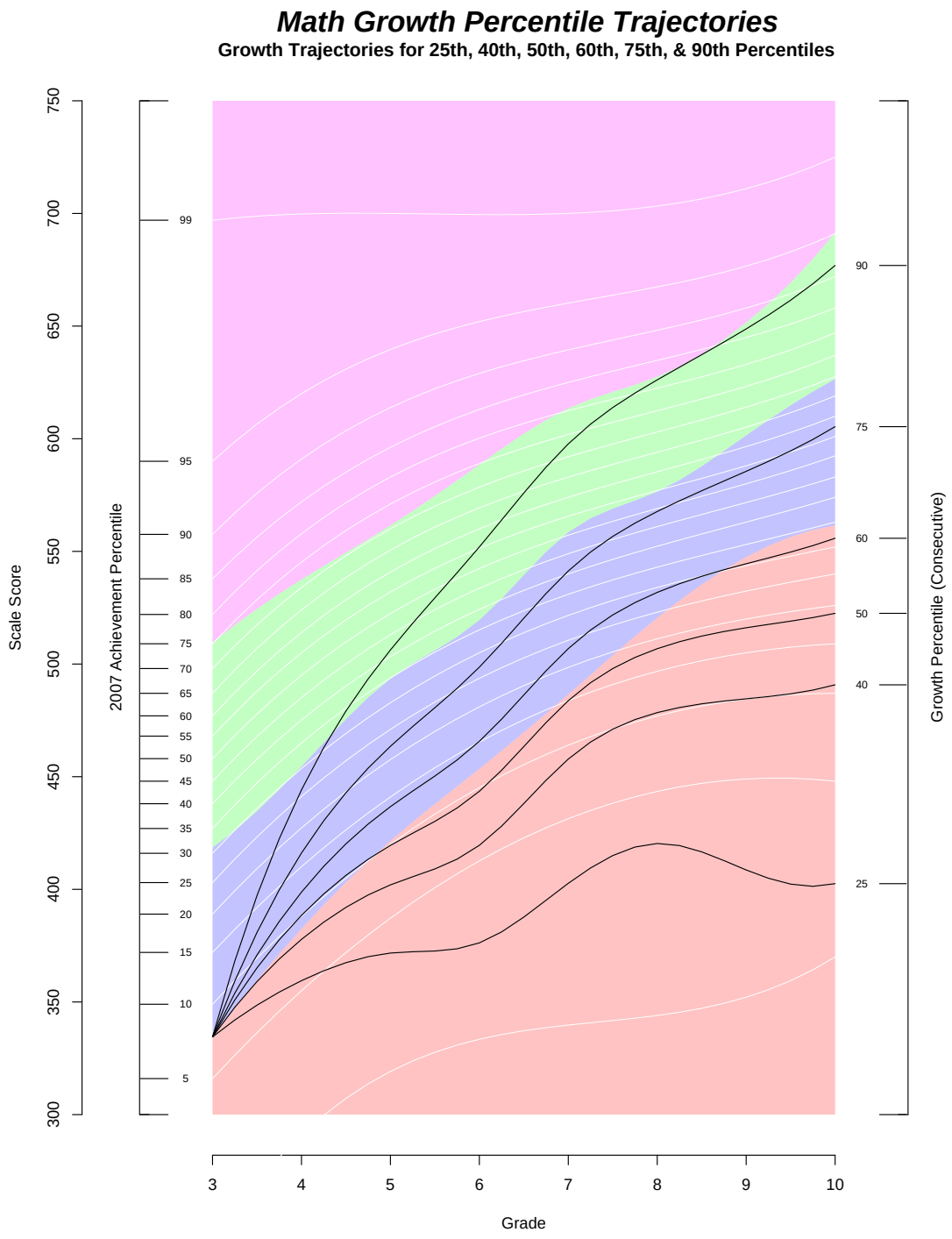

Figure 4: Growth chart depicting future math achievement conditional upon consecutive 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the unsatisfactory/partially proficient cutpoint

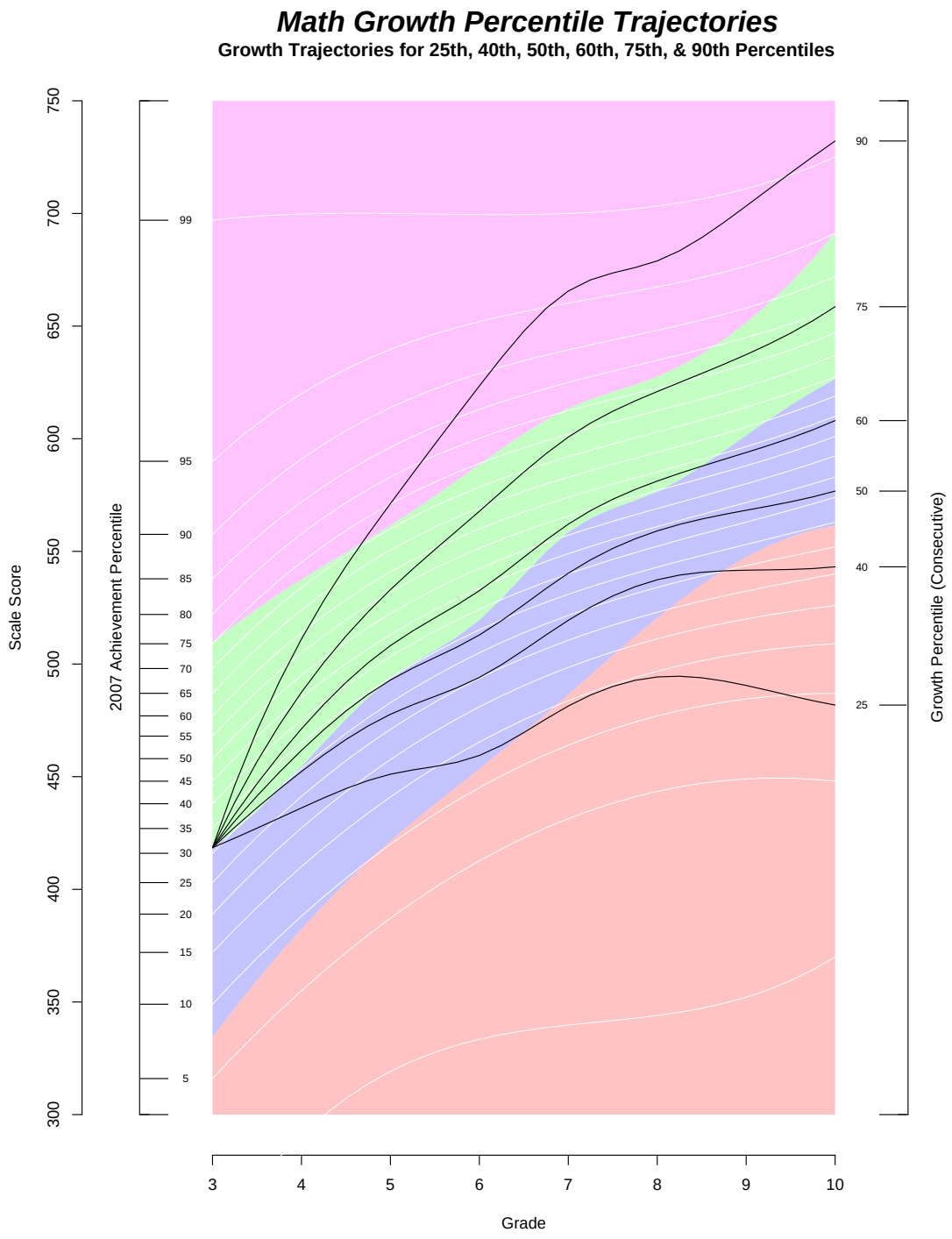

Figure 5: Growth chart depicting future math achievement conditional upon consecutive 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the partially proficient/proficient cutpoint

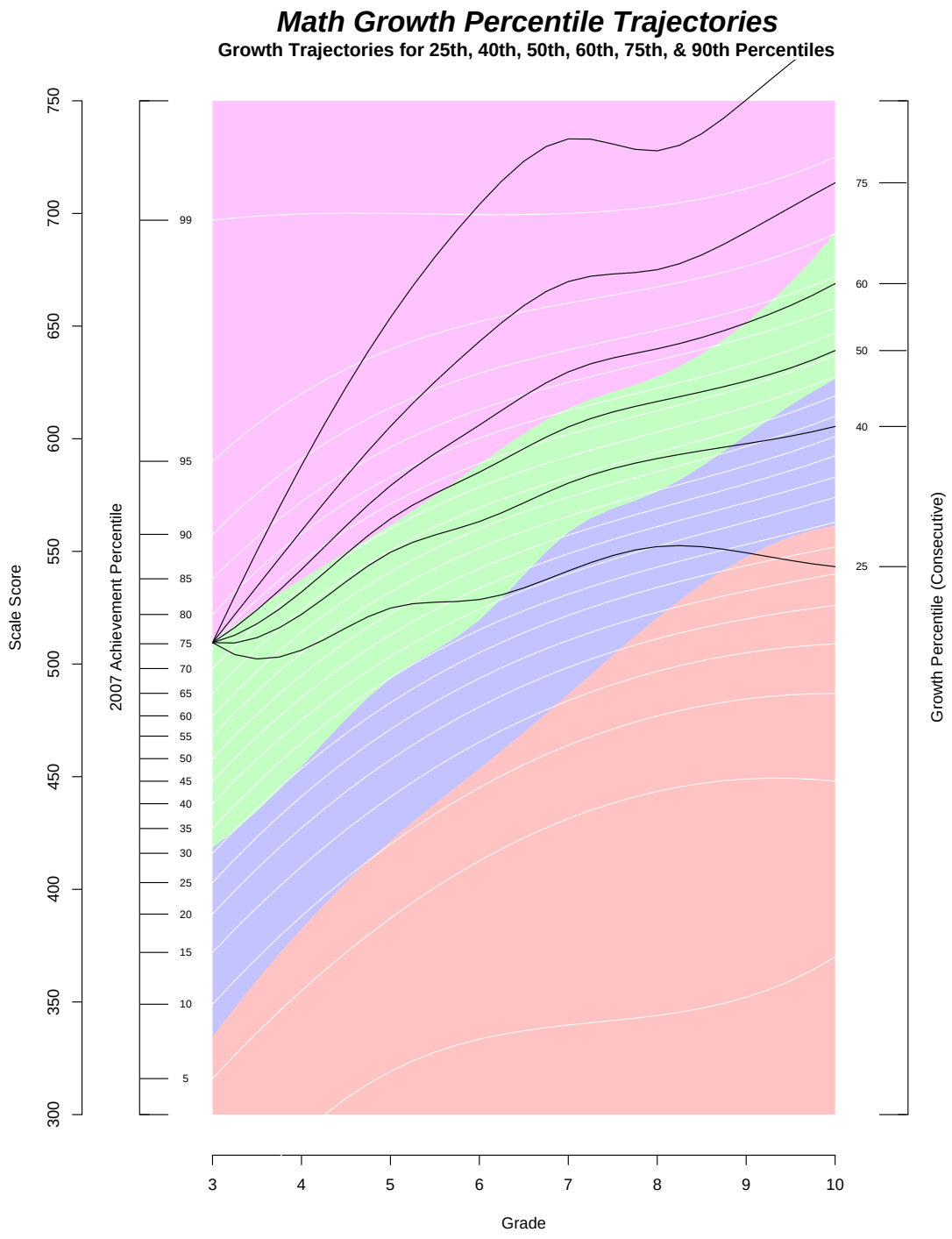

Figure 6: Growth chart depicting future math achievement conditional upon consecutive 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the proficient/advanced cutpoint

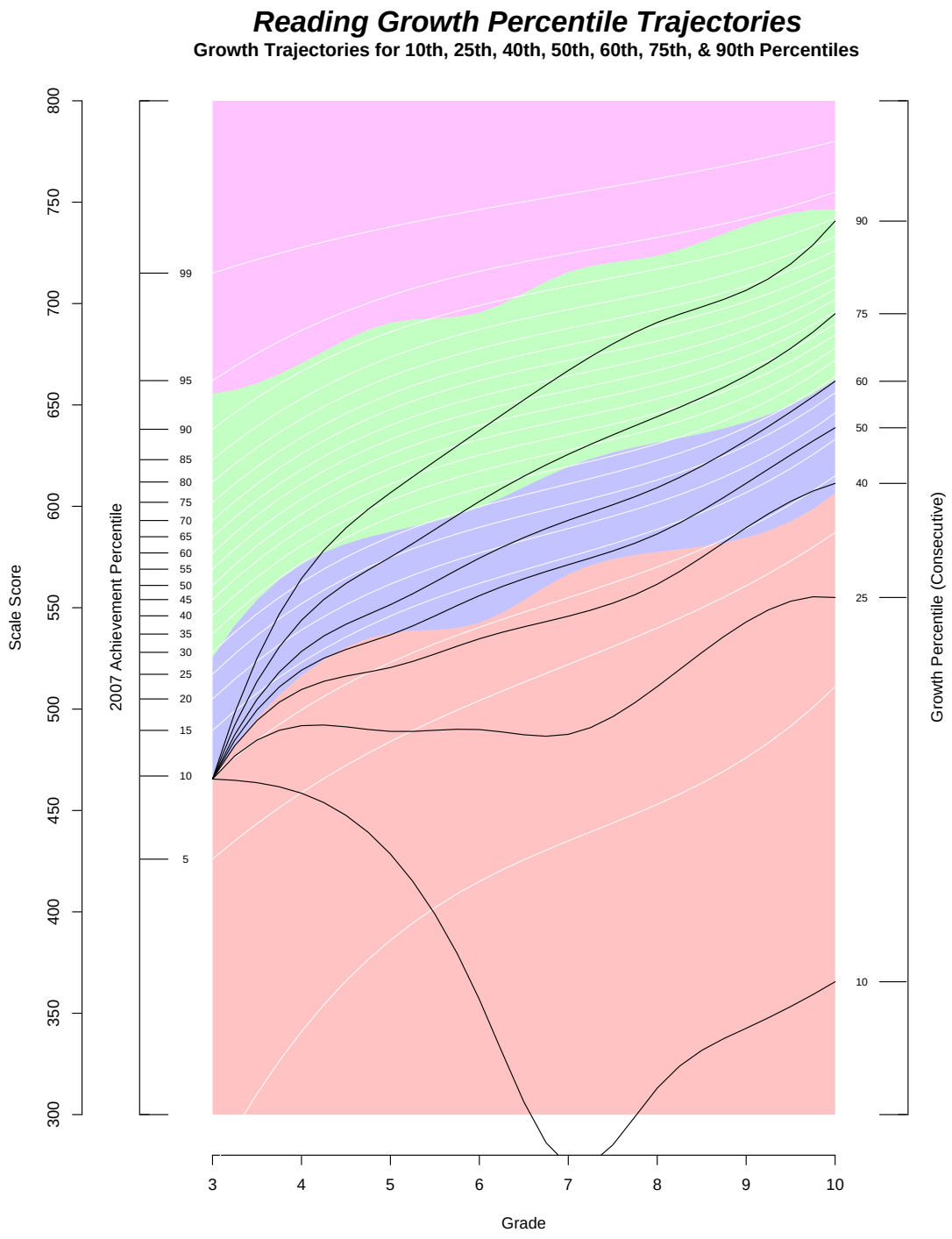

Figure 7: Growth chart depicting future reading achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the unsatisfactory/-partially proficient cutpoint

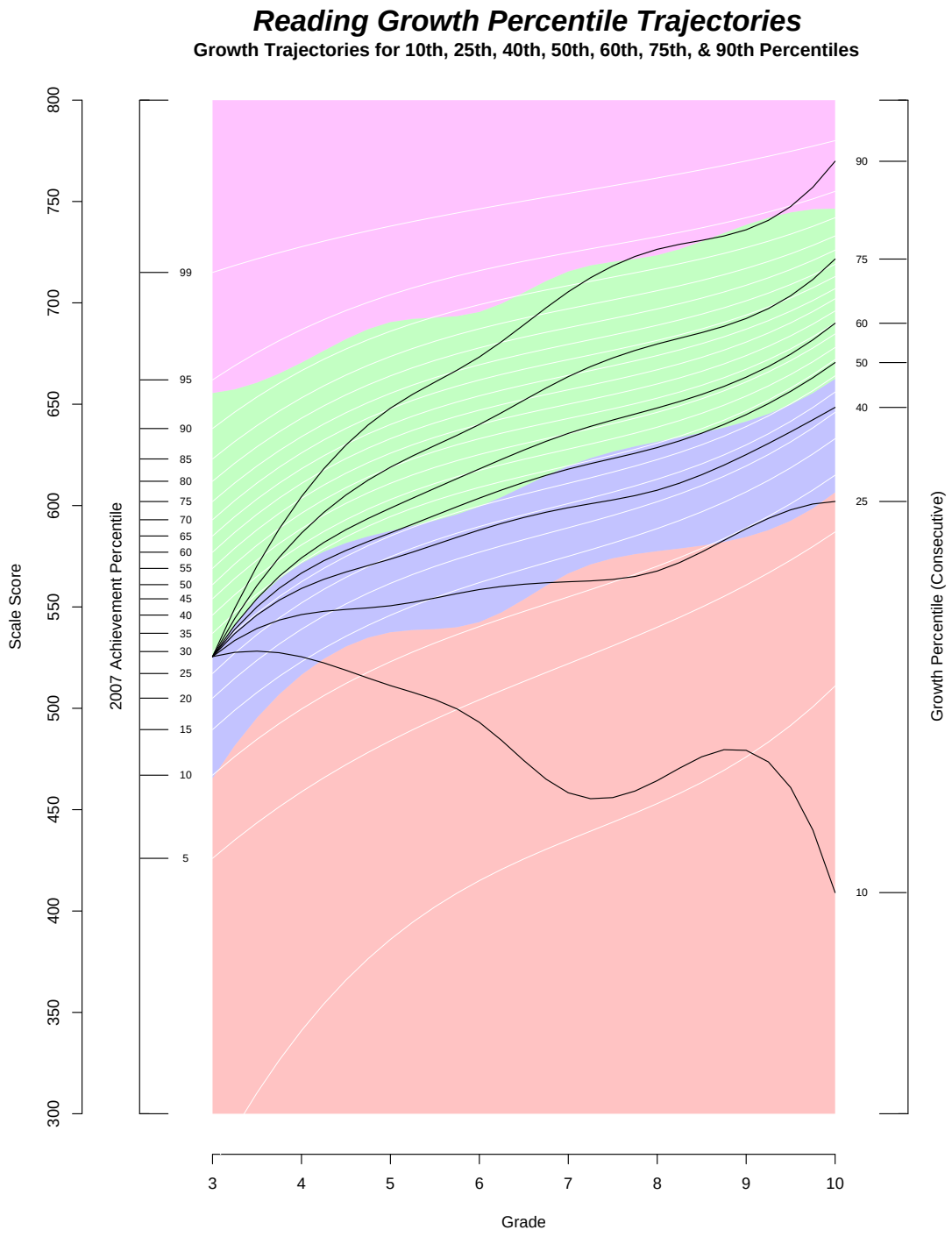

Figure 8: Growth chart depicting future reading achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the partially proficient/proficient cutpoint

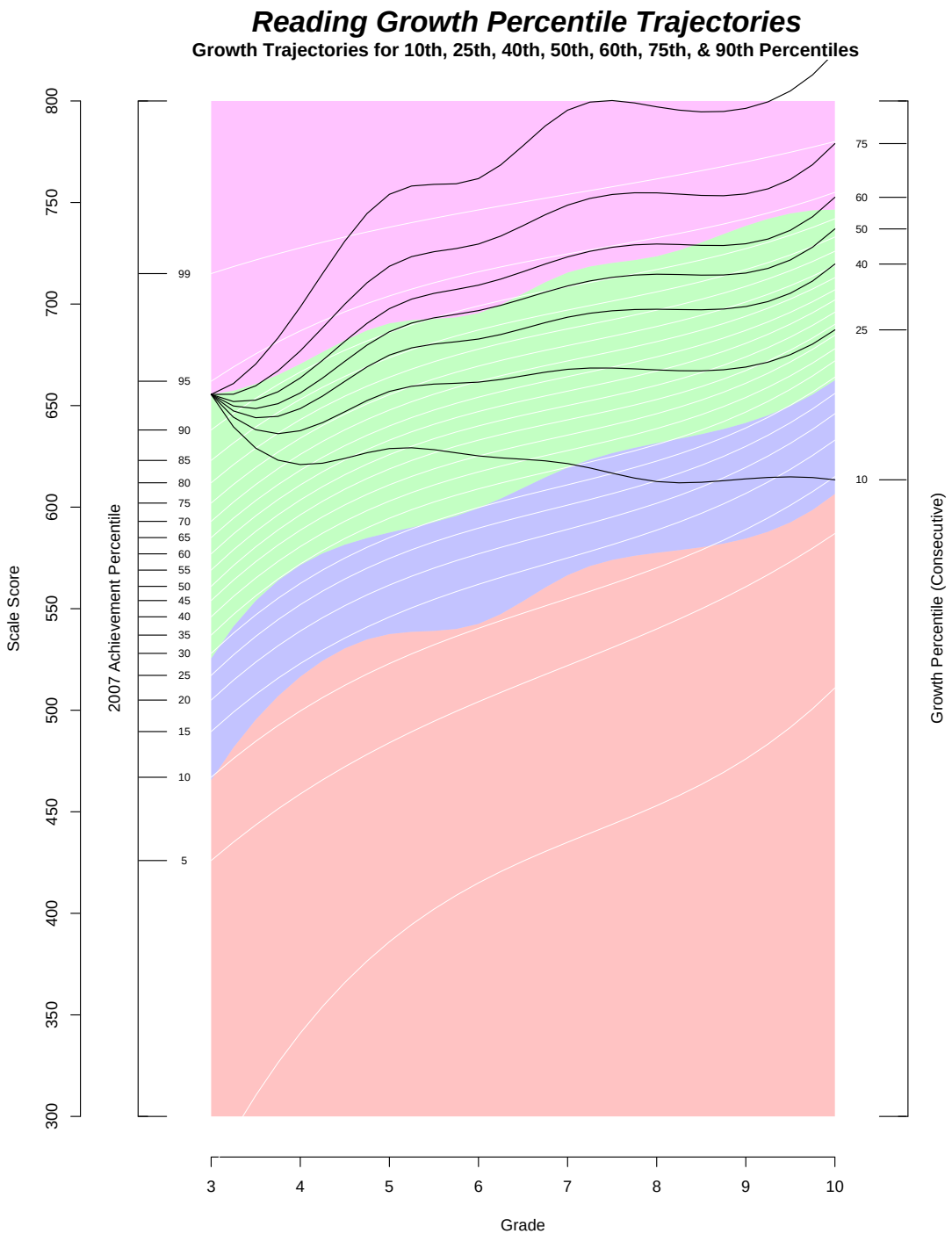

Figure 9: Growth chart depicting future reading achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the proficient/advanced cutpoint

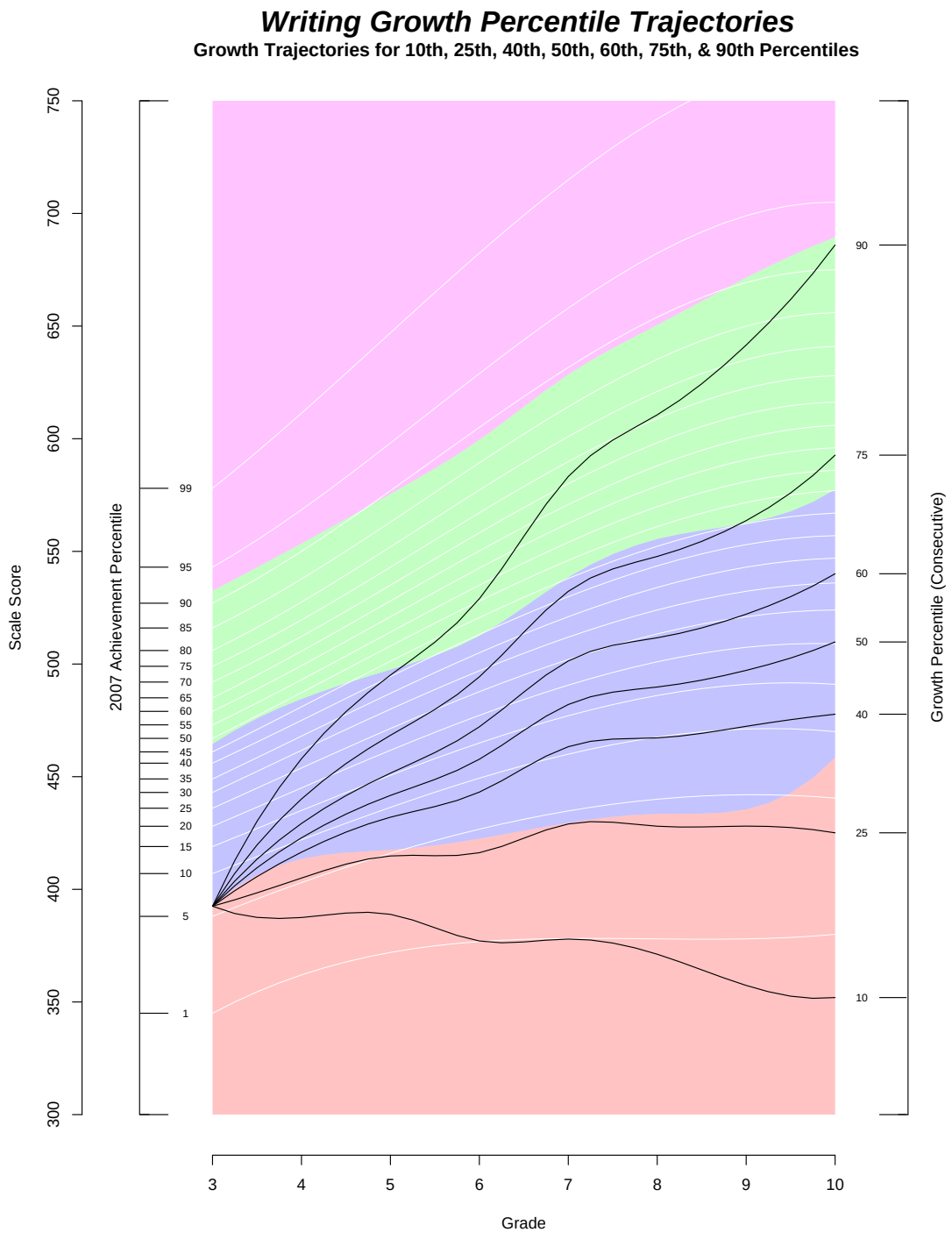

Figure 10: Growth chart depicting future writing achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the unsatisfactory/-partially proficient cutpoint

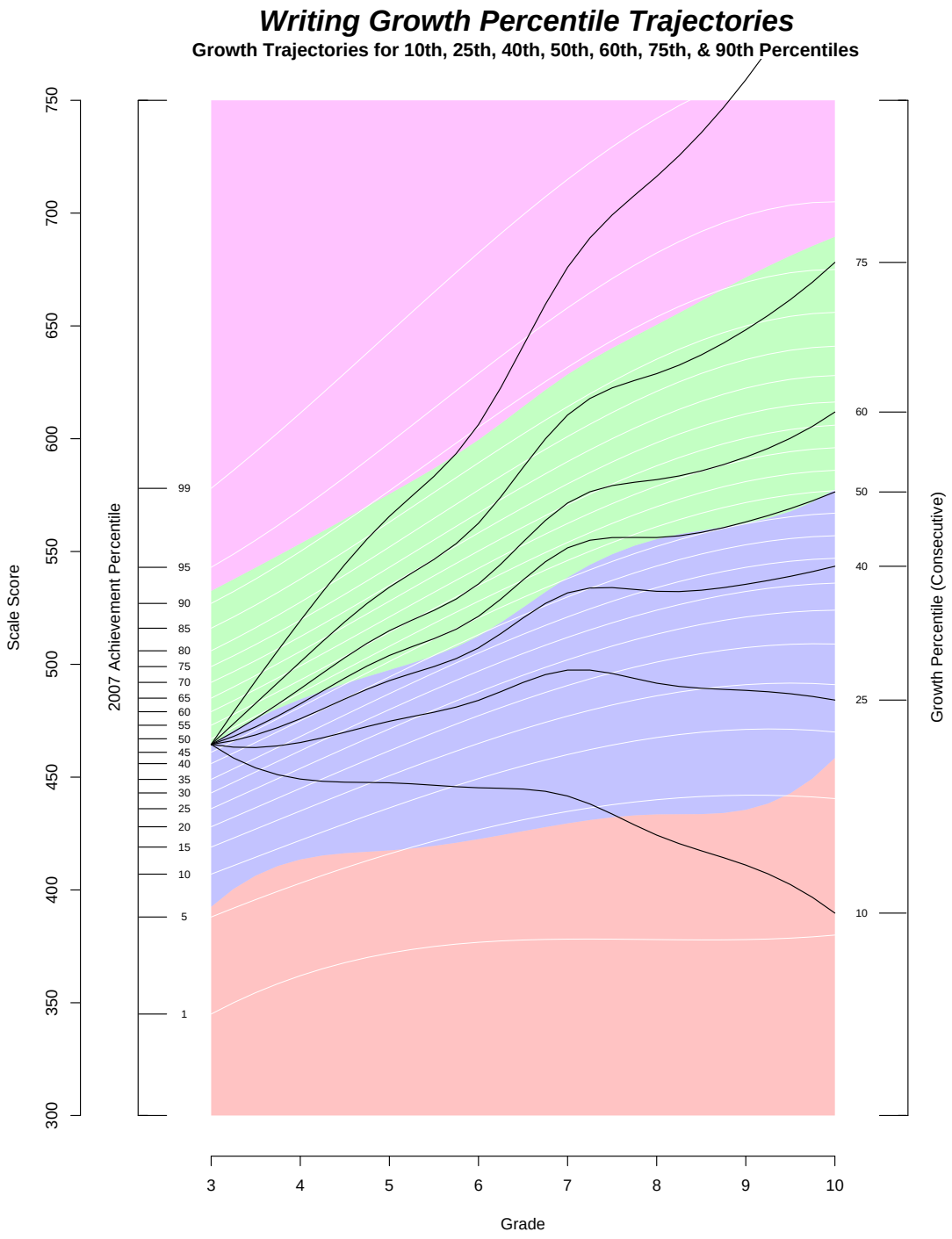

Figure 11: Growth chart depicting future writing achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the partially proficient/proficient cutpoint

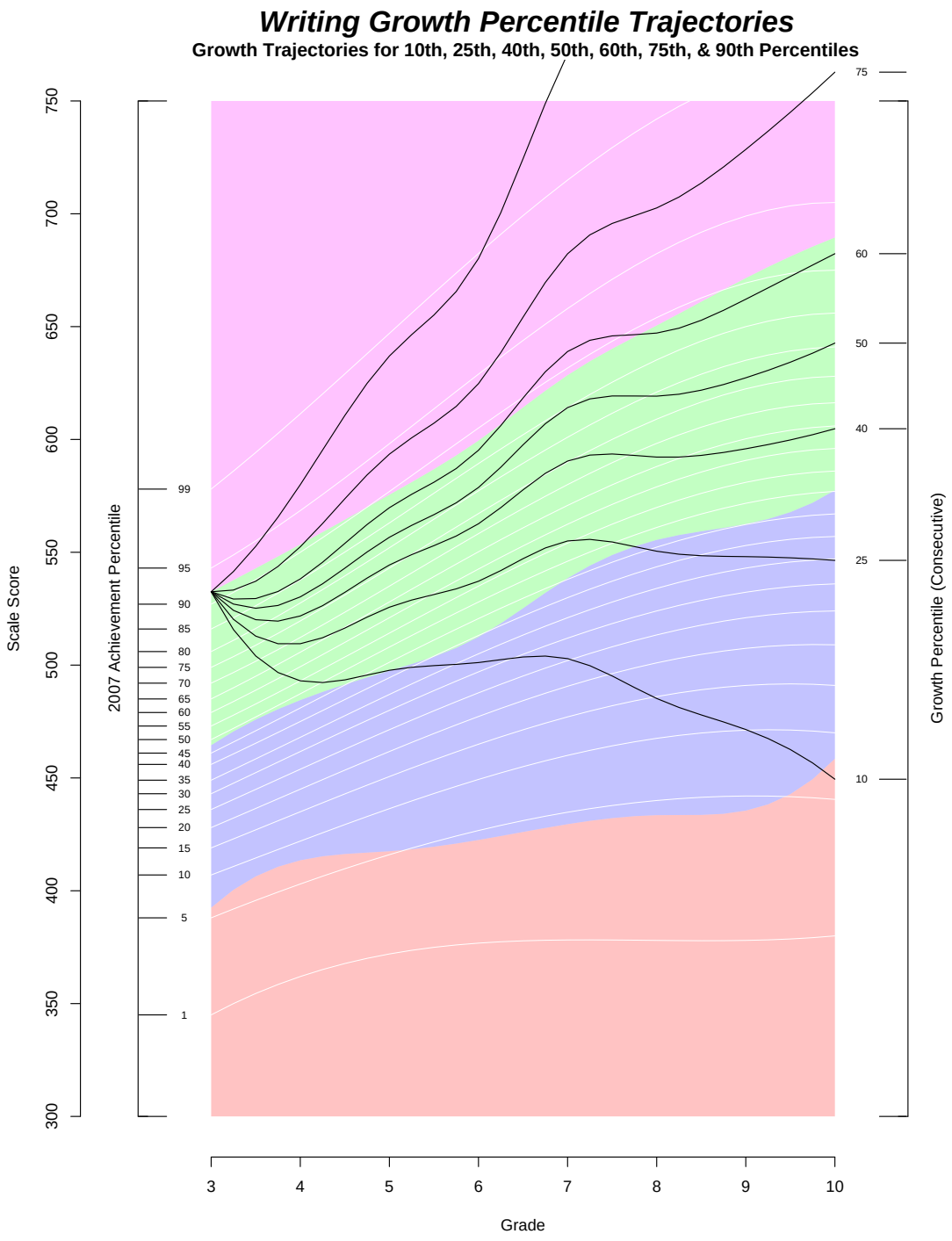

Figure 12: Growth chart depicting future writing achievement conditional upon consecutive 10th, 25th, 40th, 50th, 60th, 75th, and 90th percentile growth for a student beginning the third grade at the proficient/advanced cutpoint