Appendix B to the Utah Growth Model Report
An Overview of the SGP Methodology

Damian W. Betebenner
Adam R. VanIwaarden
National Center for the Improvement of Educational Assessment (NCIEA)

May 2015

1 Introduction - Why Student Growth?

Accountability systems constructed according to federal adequate yearly progress (AYP) requirements currently rely upon annual "snap-shots" 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 and the Race to the top competitive grants program were met with great enthusiasm by states (Spellings, 2005).

Following these use cases, 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 - that is, to disentangle current *aggregate* level achievement from effectiveness (Ballou, Sanders, & Wright, 2004; Braun, 2005; Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004). Such analyses, often called *value-added* analyses, attempt to estimate the teacher or school contribution to student achievement. This contribution, called the *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 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 questions of interest and the statistical model employed to answer those questions. Along these lines, (Harris, 2007) distinguishes value-added for program evaluation (VAM-P) and value-added for accountability (VAM-A) - conceptualizing accountability as a difficult type of program evaluation. Indeed, the current climate of high-stakes, test-based accountability has blurred the lines between program evaluation and accountability. This, combined with the emphasis of value-added models toward causal claims regarding 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 are 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 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). 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.

In the following paragraphs, student growth percentiles and percentile growth projections/-trajectories are introduced as a means of understanding student growth in both norm-referenced and criterion referenced ways. With these values calculated we show how growth data can be utilized in both a norm- and in a criterion-referenced manner to inform discussion about education quality. We assert that the establishment of a norm-referenced basis for student growth eliminates a number of the problems of incorporating growth into accountability systems providing needed insight to various stakeholders by addressing the basic question of how much a student has progressed (Betebenner, 2008; D. W. Betebenner, 2009).

2 Student Growth Percentiles

It is a common misconception that to quantify student progress 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 concepts necessary to fully understand student growth. Growth, fundamentally, requires change to be examined for a single construct like math achievement across time - *growth in what?*

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 towards but can never meet.[1]

> 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 measure 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. In this situation, parents are not interested in an absolute criterion of growth, but instead in a norm-referenced 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 one to diagnose how (a)typical such an increase is.

Given this reality in the examination of change where scales of measurement are perfect, we argue that it is unreasonable to think that in education, where scales are at best quasi-interval (Lord, 1975; Yen, 1986) one can/should examine growth differently.

Going further, suppose that 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 to a parent's question such as, "How much did my child progress?", would be a number of scale score points - an answer that would leave most parents confused wondering whether the number of points is good or bad. As in pediatrics, the search for a description regarding changes in achievement over time (i.e., growth) is best served by considering a norm-referenced quantification of student growth - *a student growth percentile* (Betebenner, 2008; D. W. Betebenner, 2009).

A student's growth percentile (SGP) describes how (a)typical a student's growth is by examining his/her current achievement relative to his/her *academic peers* - those students beginning at the same place. 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 open the infinite 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 - his/her academic peers. If the student's current year score exceeded the scores of most of his/her academic peers, in a norm-referenced

[1] The scales on which students are measured are often assumed to possess properties similar to height and weight but they don't. Specifically, scales are assumed to be interval where it is assumed that a difference of 100 points at the lower end of the scale refers to the same difference in ability/achievement as 100 points at the upper end of the scale. (See Lord, 1975; and Yen, 1986 for more detail on the interval scaling in educational measurement.)

sense they have done as well. If the student's current year score was less than the scores of his/her academic peers, in a norm-referenced sense they have not done as well.

The four panels of Figure B.1. 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 2011 and 2012 scores for all students in the state yields a bivariate (two variable) distribution. The higher the distribution, the more frequent the pair of scores.

- **Upper Right Panel** Taking account of prior achievement (i.e., conditioning upon prior achievement) fixes the value of the 2011 scale score (in this case at approximately 460) 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 2012 test assuming a 2011 score of 460. This distribution is indicated by the solid red slice of the distribution.

- **Lower Right Panel** The conditional distribution provides the context against which a student's 2012 achievement can be examined and provides the basis for a norm-referenced comparison. 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 500 on the 2012 test. Within the conditional distribution, the value of 500 lies at the 75^{th} percentile. Thus the student's progress from 460 in 2011 to 500 in 2012 met or exceeded that of 75 percent of students starting from the same place. It is important to note that qualifying a student growth percentile as "adequate", "good", or "enough" is a standard setting procedure that requires stakeholders to examine a student's growth *vis-a-vis* external criteria such as performance standards/levels.

Fig. B.1: Depiction of the distribution associated with 2011 and 2012 student scale scores together with the conditional distribution and associated growth percentile.

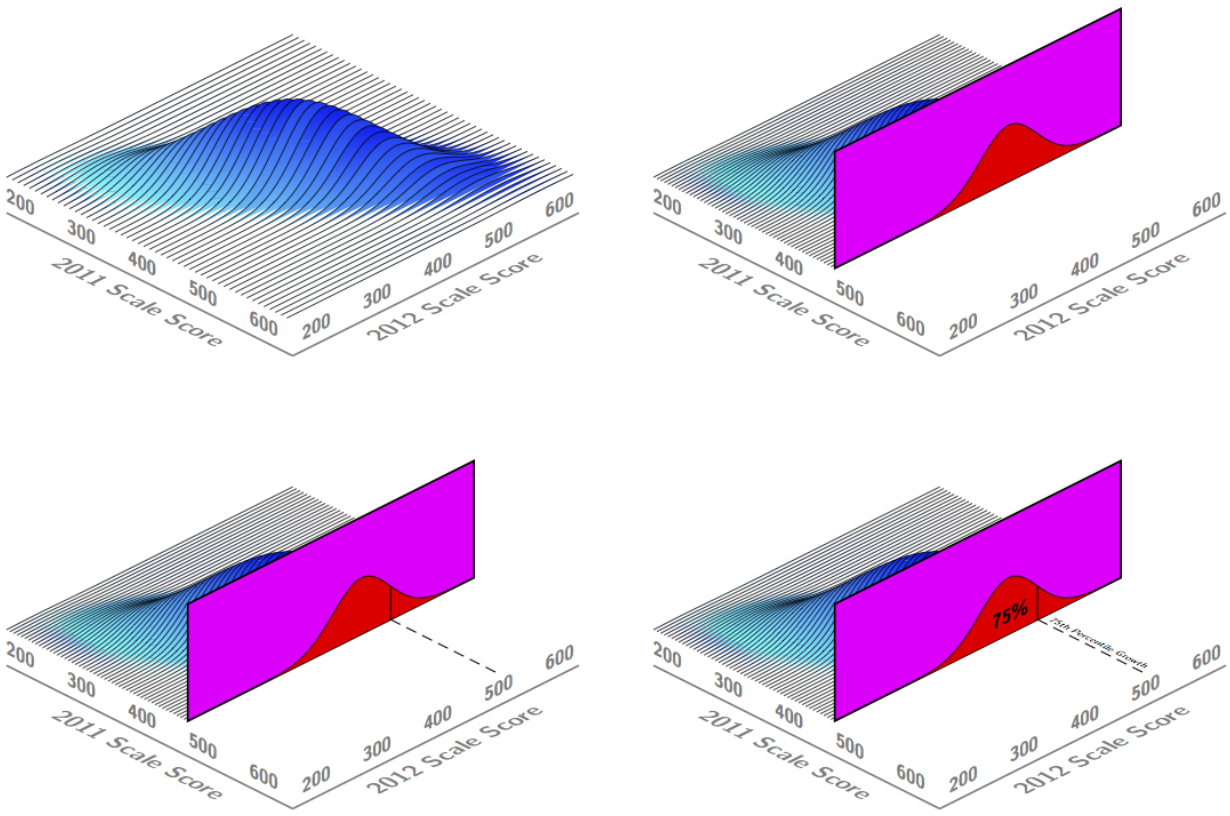


Figure B.1 also serves to illustrate the relationship between the state's assessment scale and student growth percentiles. The scale depicted in the panels of Figure B.1 is not vertical. Thus the comparisons or subtraction of scale scores for individual students is not supported. However, were such a scale in place, the figure would not change. With or without a vertical scale, the conditional distribution can be constructed.

In situations where a vertical scale exists, the increase/decrease in scale score points can be calculated and the growth percentile can be understood alongside this change. For example, were the scales presented in Figure B.1 vertical, then one can calculate that the student grew 40 points (from 460 to 500) between 2011 and 2012. This 40 points represents the absolute magnitude of change. Quantifying the magnitude of change is scale dependent. For example, different vertical achievement scales in 2011 and 2012 would yield different annual scale score increases: A scale score increase of 40 could be changed to a scale score increase of 10 using a simple transformation of the vertical scale on which all the students are measured. However, relative to other students, their growth has not changed - their growth percentile is invariant to scale transformations common in educational assessment. Student growth percentiles norm-referencedly situate achievement change bypassing questions associated with the magnitude of change, and directing attention toward relative standing which, we would assert, is what stakeholders are most interested in.

To fully understand how many states intend to use growth percentiles to make determinations about whether a student's growth is sufficient, the next section details specifics of how student growth percentiles are calculated. These calculations are subsequently used to calculate percentile growth projections/trajectories that are used to establish how much growth it will take for each student to reach his/her achievement targets.

3 SGP Calculation

Quantile regression is used to establish curvilinear functional relationships between the cohort's prior scores and their current scores. Specifically, for each grade by subject cohort, quantile regression is used to establish 100 (1 for each percentile) curvilinear functional relationships between the students prior score(s) and their current score. For example, consider 7^{th} graders in 2014. Their grade 3, grade 4, grade 5, and grade 6 prior scores are used to describe the current year grade 7 score distribution.[2] The result of these 100 separate analyses is a single coefficient matrix that can be employed as a look-up table relating prior student achievement to current achievement for each percentile. Using the coefficient matrix, one can plug in *any* grade 3, 4, 5, and 6 prior score combination to the functional relationship to get the percentile cutpoints for grade 7 conditional achievement distribution associated with that prior score combination. These cutpoints are the percentiles of the conditional distribution associated with the individual's prior achievement. Consider a student with the following mathematics scores:

Table 1: Scale scores for a hypothetical student across 5 years in mathematics.

Grade 3/2010	Grade 4/2011	Grade 5/2012	Grade 6/2013	Grade 7/2014
819	818	822	834	836

Using the coefficient matrix derived from the quantile regression analyses based upon grade 3, 4, 5, and 6 scale scores as independent variables and the grade 7 scale score as the dependent variable together with this student's vector of grade 3, 4, 5, and 6 grade scale scores provides the scale score percentile cutpoints associated with the grade 7 conditional distribution for these prior scores.

Table 2: Percentile cutscores for grade 7 mathematics based upon the grade 3, 4, 5, and 6 mathematics scale scores given in Table 1.

1st	2nd	3rd	...	10th	...	25th	...	50th	51th	...	75th	...	90th	...	99th
804.8	814.9	819.9	...	825.9	...	830.8	...	835.5	836.3	...	868.9	...	887.1	...	909.8

The percentile cutscores for 7^{th} grade mathematics in Table FALSE are used with the student's *actual* grade 7 mathematics scale score to establish his/her growth percentile. In this case, the student's grade 7 scale score of 836 lies above the 50^{th} percentile cut and below the 51^{st} percentile cut, yielding a growth percentile of 50. Thus, the progress demonstrated by this student between grade 6 and grade 7 exceeded that of 50 percent of his/her academic peers - those students with the same achievement history. States can qualify student growth by defining ranges of growth percentiles. For example, the Utah Growth Model designates growth

[2]For the mathematical details underlying the use of quantile regression in calculating student growth percentiles, see the *SGP Estimation* section

percentiles between 35 and 65 as being *typical*. Using Table FALSE, another student with the exact same grade 3, 4, 5, and 6 prior scores but with a grade 7 scale score of 804, would have a growth percentile of 1, which is designated as *low*.

This example provides the basis for beginning to understand how growth percentiles in the SGP Methodology are used to determine whether a student's growth is *(in)adequate*. Suppose that in grade 6 a one-year (i.e., 7^{th} grade) achievement goal/target of proficiency was established for the student. Using the lowest proficient scale score for 7^{th} grade mathematics, this target corresponds to a scale score of 900. Based upon the results of the growth percentile analysis, this one year target corresponds to 95^{th} percentile growth. Their growth, obviously, is less than this and the student has not met this individualized growth standard.

4 SGP 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 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 characterize the student's 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). 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

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)) = \begin{cases} u \cdot \tau & u \geq 0 \\ u \cdot (\tau - 1) & u < 0. \end{cases} \tag{2}$$

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 τ^{th} sample quantile associated with Y. By conditioning on a covariate x, the τ^{th} conditional quantile function is given by

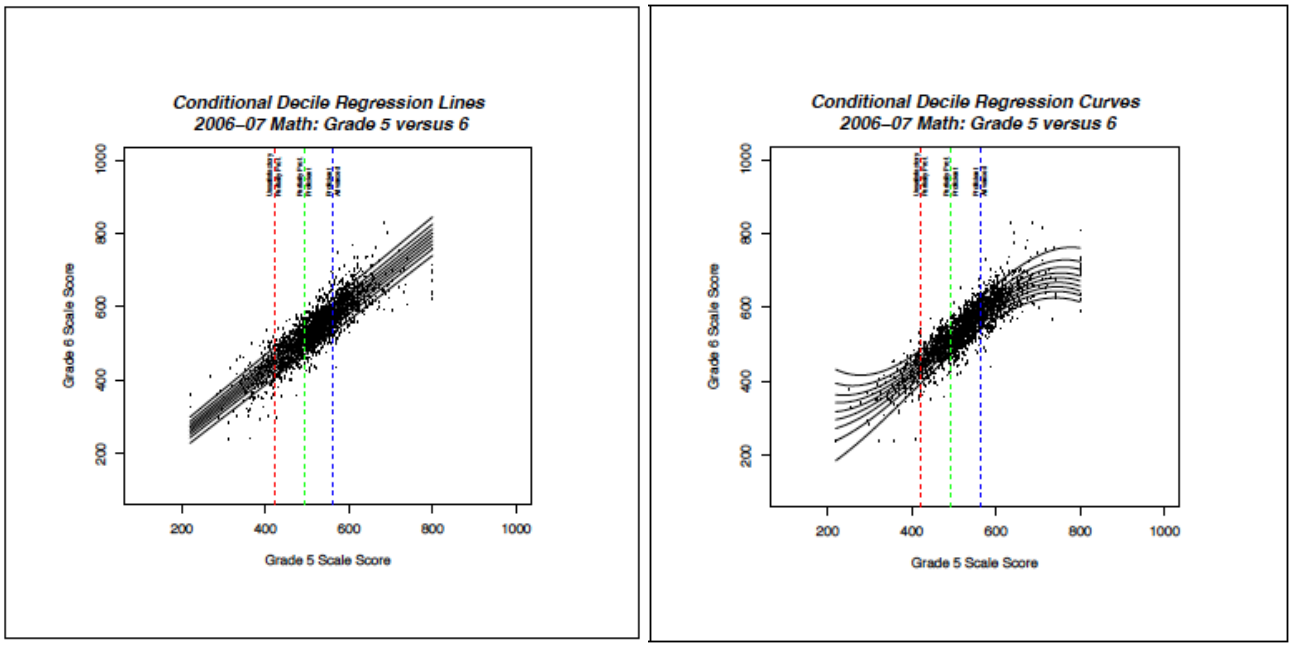
$$Q_y(\tau|x) = arg\ min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta). \tag{3}$$

In particular, if $\tau = 0.5$, 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), and are simple to implement in available software.

Figure B.2 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.

Fig. B.2: Linear and B-spline conditional deciles based upon bivariate math data, grades 5 and 6.



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

Calculation of student growth percentiles is performed using `R` (R Development Core Team, 2015), a language and environment for statistical computing, with `SGP` package (Betebenner, VanIwaarden, Domingue, & Shang, 2014). Other possible software (untested with regard to student growth percentiles) with quantile regression capability include SAS and Stata. Estimation of cohort referenced 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

$$Q_{Y_t}(\tau|Y_{t-1}, \ldots, Y_1) = \sum_{j=1}^{t-1} \sum_{i=1}^{3} \phi_{ij}(Y_j)\beta_{ij}(\tau), \tag{4}$$

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 B.2 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 B-spline basis functions require the selection of boundary and interior knots. Boundary knots are end points outside of the scale score distribution that anchor the B-spline basis. These are generally selected by extending the range of scale scores by 10%. That is, they are defined as lying 10% below the lowest obtainable (or observed) scale score (LOSS) and 10% above the highest obtainable scale score (HOSS). The interior knots are the *internal* breakpoints that define the spline.

The default choice in the `SGP` package (Betebenner et al., 2014) is to select the 20^{th}, 40^{th}, 60^{th} and 80^{th} quantiles of the observed scale score distribution. In general the knots and boundaries are computed using a distribution from several years of compiled test data (i.e. multiple cohorts) so that any irregularities in a single year are smoothed out. Subsequent annual analyses then use these same knots and boundaries as well. All defaults were used to compile the knots and boundaries for Utah from the CRT tests. New knots and boundaries were required beginning with the 2015 SGP analyses when SAGE assessments were first used as dependent variables in the quantile regressions.

Finally, it should be noted that the independent estimation of the regression functions can potentially result in the crossing of the quantile functions. This occurs near the extremes of the distributions and is potentially more likely to occur given the use of non-linear functions. The result of allowing the quantile functions to cross in this manner would be *lower* percentile estimations of growth for *higher* observed scale scores at the extremes (give all else equal in prior scores) and vice versa. In order to deal with these contradictory estimates, quantile regression results are isotonized to prevent quantile crossing following the methods derived by Chernozhukov, Fernandez-Val and Glichon (2010).

5 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 norm-referenced context for Linn's existence proof (Linn, 2003) 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 occasion 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 norm-referenced 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 norm-referencedly 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.[4]

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

$$Q_h(Y)|X(\tau|X) = h(Q_Y|X(\tau|X)). \tag{5}$$

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.

[4] 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.

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. (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). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2014). *SGP: An r package for the calculation and visualization of student growth percentiles & percentile growth trajectories.* Retrieved from https://github.com/CenterForAssessment/SGP

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

Chernozhukov, V., Fernndez-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125. Wiley Online Library.

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.* Princeton, NJ: Educational Testing Service.

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

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations.* 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.

Lord, F. M. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika*, *20*, 299–326.

R Development Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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). *Secretary Spellings Announces Growth Model Pilot.* Press Release, U.S. Department of Education.

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

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

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

New York: Springer.