Making Inferences about Growth and Value-Added:
Design Issues for the PARCC Consortium

Derek C. Briggs
University of Colorado, Boulder

December 28, 2011

A White Paper Commissioned by the PARCC Consortium.

ABSTRACT

There is often confusion about distinctions between growth models and value-added mdoels. The first half of this paper attempts to dispel some of these confusions by clarifying terminology and illustrating by example how the results from a large-scale assesment can and will be used to make inferences about student growth and the value-added attributable to teachers or schools. Two key differences between growth models and value-added models are discussed: the unit of analysis (growth models focus first and foremost on students; value-added models focus on teachers or schools) and inferential intent (growth models are primarily descriptive, value-added models are meant to support causal inferences). The point is made that all growth models can be used to make value-added inferences, but value-added models almost never lead to student-specific inferences about growth. The focus of the second half of this paper is on design issues that will need to be considered by the PARCC consortium such that test scores can be used for either growth or value-added inferences. It is shown that vertically scaled test scores are *not* prerequisite for value-added modeling. Vertical scales are most desirable in support of student-level growth interpretations, but certain conditions must be met before their creation would be defensible. In particular, the case is made that vertical scales will be most compatible with a learning progression perspective on construct and item development. The second half of the paper also discusses design factors that would minimize the role that measurement error will play in distorting inferences about growth and value-added. Finally, the paper concludes with the some recommendations for the PARCC consortia.

Introduction

One of the priority purposes that has been expressed for the PARCC assessments is that they should “provide information, including measures of growth, that supports various forms of accountability including teacher effectiveness. (PARCC Assessment System: Requirements and Constraints, June 17, 2011)” I have written this paper with two primary objectives in mind. The first objective is to attempt to clarify, or at least bring to the surface, some understandable confusion over terminology. For example, what is a growth model? How is a growth model different from a value-added model? Do statements about growth or value-added require a vertical scale? It seems important for the state members of PARCC to come to some common understandings when answering such questions. The second objective is to discuss design factors for the PARCC assessments that will have a bearing on the inferences that can be supported about growth and value-added.

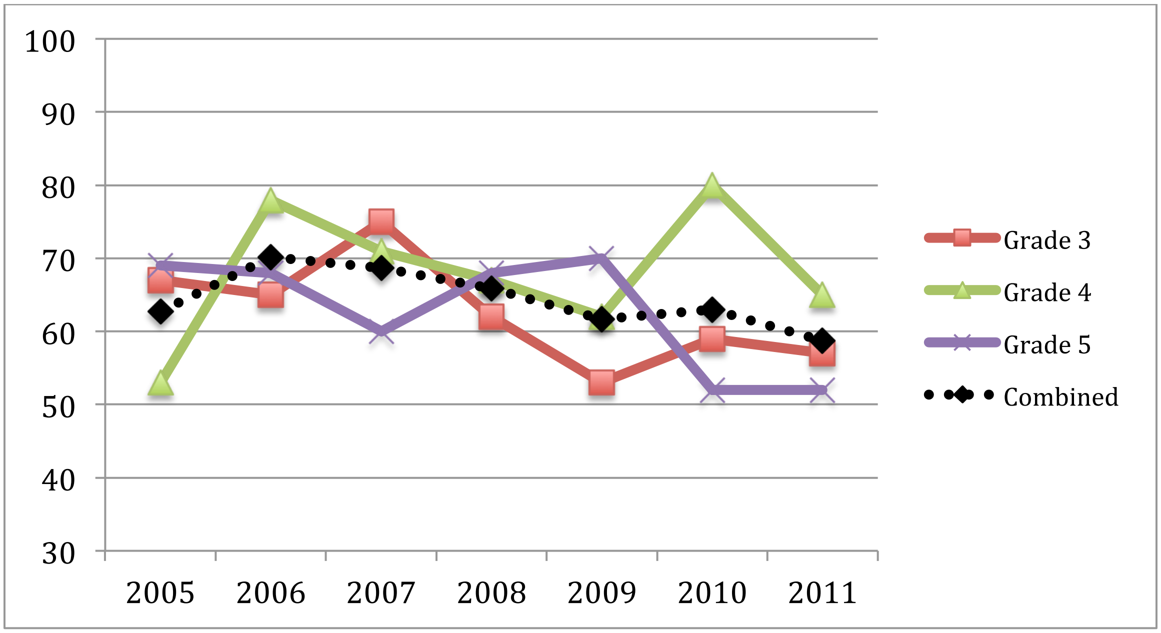
To give this paper some added bite at the outset, and to further motivate animated discussion, I have chosen to illustrate a number of salient issues and distinctions in the context of the real-world example of an elementary school in Colorado. This school was recently sanctioned under the auspices of No Child Left Behind because over the course of two years its students had not met performance targets on the Colorado Student Assessment Program (CSAP) tests. In the future, one of the most identifiable features of No Child Left Behind is likely to change upon reauthorization: namely, inferences about school performance will shift from whether or not students have reached some minimum level of proficiency to whether they have obtained some minimum level of college and career readiness. Nonetheless, the fundamental issue of how scores from a large-scale assessment can best be used to make inferences about a student’s academic preparation and the quality with which this is being provided in public school settings will remain the same. What would need to be said about the PARCC tests relative to the CSAP tests such that the use of scores for educational accountability purposes had greater validity? Furthermore, what can growth or value-added models bring to the table (that has not been there in the past) and what will they require of the PARCC assessments?

Creekside Elementary School and Growth

On August 3, 2011, approximately one week before the start of the school year, the parents of students attending Creekside Elementary School received a letter informing them that, because Creekside had failed the provisions required to demonstrate adequate yearly progress for two years in a row, they were eligible to have their children attend a different school in the Boulder Valley School District. If such a choice were to be made, transportation would be provided at the district's expense. The rest of the letter detailed the steps that Creekside was and would continue to take to improve classroom teaching and learning, with the implicit message being "Don't give up on us yet!" Empirically, there is little question that the average academic achievement of students at Creekside leaves much to be desired. Among students tested in grades 3-5 in math, reading and writing on the CSAP test, only 59%, 57% and 47% of students scored at a level that would be classified as either "proficient" or "advanced" according to Colorado's performance standards. Across the entire school district these numbers were considerably higher (70%, 80% and 69%). On the other hand, it is also true that Creekside faces certain challenges with the student population it is educating that the higher-achieving schools in Boulder County do not. For example, the proportions of Hispanic students (many of whom are likely to be English Language Learners) and students eligible for free and reduced lunch services at Creekside (40% and 33%) is more than twice as large as the proportions across the district (18% and 17%). Some of the current interest in evaluating test scores for evidence of growth is that this should make it possible to better distinguish between schools where students may be low-achieving but making laudable progress, and schools with students who are low-achieving and making either no progress or progress that is negligible. Would such an approach make a difference when casting judgment upon the quality of education at Creekside? In this section I do my best to present the relevant data and let the reader be the judge. To keep the presentation manageable, I focus attention solely on CSAP test performance in mathematics.

Cohort to Cohort Change

Figure 1. *Cohort to Cohort Change in Percent Proficient or Advanced at Creekside Elementary in Mathematics, 2005-2011*



Braun et al (2010) describe a "cohort to cohort change model" as one that compares achievement status at two or more points in time, but not for the same students. As an example of this, Figure 1 plots the grade 3, 4 and 5 percentages of students classified as either "proficient" or "advanced" on the CSAP math exams from 2005 to 2011. Also included in this plot is the trend line for the average percentage across all three grades. Averaging across grades has the effect of greatly reducing the year to year variability in the percentages. This happens for two reasons. First, to the extent that some of the variability from year to year can be attributed to either measurement error or sampling error in a way that is analogous to sampling theory, an increase in the number of students used to compute a school-level average will reduce this source of variability[1]. Second, when an average is taken across grades it creates a pseudo-longitudinal data structure in which as much as 2/3 of the students may overlap across adjacent years, and

[1] The average number of grade 3, 4 and 5 students from 2005 to 2011 was 46, 40 and 39. So, for example, given two distinct grade 4 cohorts of 40 students, if just four fewer students scored above the CSAP cutpoint for proficiency, the average would swing by 10%.

this overlap greatly enhances the stability of the school-level statistics from year to year. It is, however, worth pointing out that when given data for any two year period, school administrators are likely to overinterpret grade-specific swings in one direction or the other. For example, after the 2009-10 school year, 59%, 80%, and 52% of grade 3, 4 and 5 Creekside students respectively had been classified as proficient or advanced in Math. On this basis, when the school's principal drafted a school improvement plan[2], these numbers were used as the baseline from which the performance of the next year's cohort would be judged. Targets were set that called for increases by 5% in the percentage of students in the 2010-11 cohorts classified as proficient or advanced in math. Empirically, it can be seen that mistaking cohort-to-cohort change for growth essentially set the school up to fail a key target on its publicly released improvement plan.

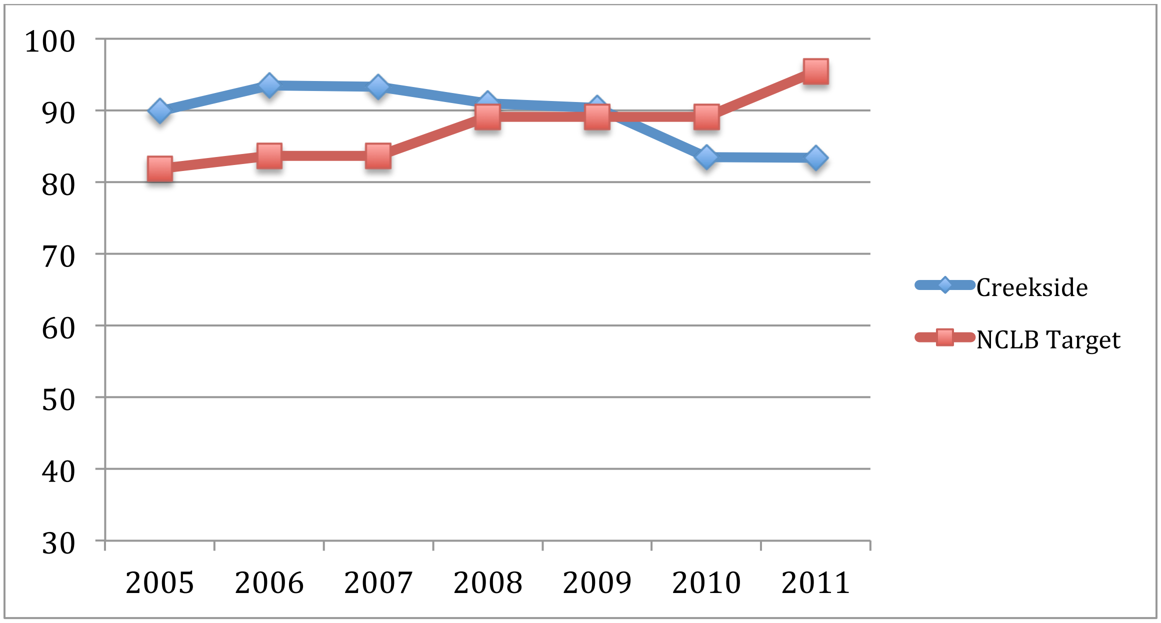
Progress to Future Target

A variant of the pseudo-longitudinal cohort-to-cohort change approach is the progress to future target model. Here the percentages of students classified as proficient or advanced in a given year are averaged across grades and then compared to a target that has been set by policymakers. In Colorado, this is a rather confusing exercise because for the purposes of compliance with No Child Left Behind, the state includes the percentage of students classified as *partially proficient* along with those classified as proficient or advanced when assessing progress toward the federal target. This has the effect of inflating upwards the values that were shown in Figure 1. Nonetheless, when observed and expected values are plotted over time as in Figure 2, a basis for Creekside's failure to make AYP in two consecutive years become evident[3].

[2] See https://cedar2.cde.state.co.us/documents/SPF2010/0480%20-%205606%20-%203%20Year.pdf

[3] The more proximal cause is far more complicated because meeting AYP is a conjunctive function of many distinct targets for different demographic subgroups, and even if a target is not met, a school may be excused through a "safe harbor" provision. See appendix Table A-1 for a flow chart that illustrates the complexity of this process in Colorado.

Figure 2. *Comparing Creekside's Progress to Standard*



Note: Values for Creekside are based on all students except those classified in the lowest performance level ("unsatisfactory"). For NCLB AYP targets, see http://www.cde.state.co.us/FedPrograms/danda/aypprof.asp

On the one hand, in Figure 2 it becomes possible to answer the policy question, "Is the progress being observed at this school good enough?" On the other hand, a longstanding criticism of No Child Left Behind has been that setting a target of 100% proficiency is unrealistic (Linn, 2003), and the example of Colorado having two different definitions of proficiency on the CSAP for state and federal accountability purposes is just one example of the way that states have attempted to put off the day of reckoning that was bound to come.

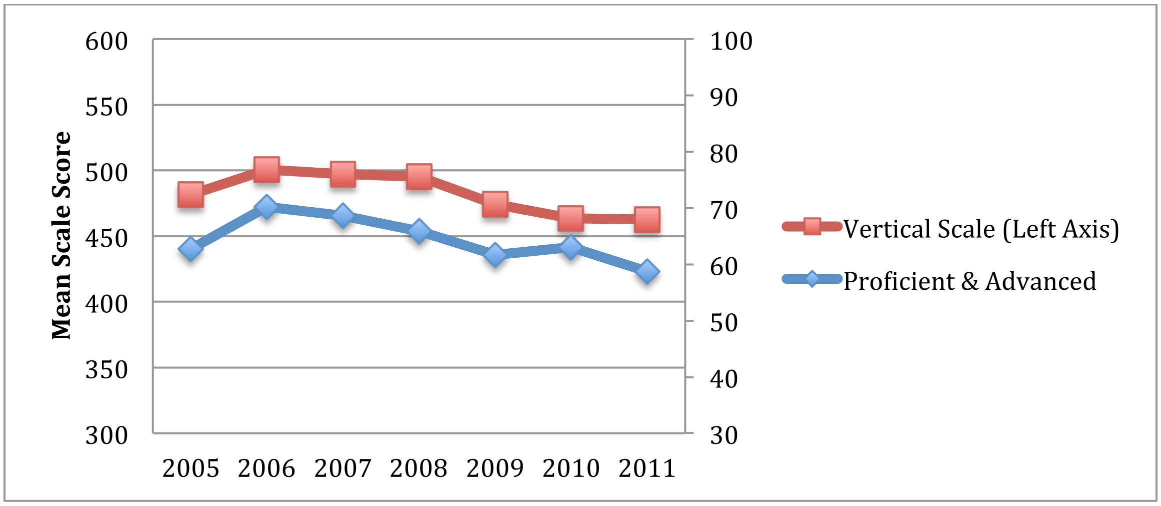
Growth Models

When the Race to the Top competition was launched by the United States Department of Education, the request for proposals included the following definition of growth as "the change in student achievement for an individual student between two or more points in time" with student achievement defined as "a student's score on one of the State's assessments…provided they are rigorous and comparable across classrooms." It follows that neither of the examples shown above would constitute a "growth model" because they are not designed to track the test scores of the *same* students from one year to the next. Though Figures 1 and 2 seem to be indicative of some worrisome trends in performance for Creekside, this interpretation is potentially confounded by the movement of students in and out of the school from year to year. If the population of students at a school is changing systematically[4] such that an increasing proportion of new students enter a given school year with academic deficits, then this could mask significant growth that might be apparent if the same students had been tracked longitudinally.

The simplest adjustment that could be made would be to continue to track trends in the proportion of students with test scores above some chosen cutpoint, as in Figures 1 and 2, but to only do some for longitudinal cohorts of students. Note that when taking such as an approach at the student level, it would only be possible to characterize growth in terms of a series of transitions from discrete classifications. A drawback to such an approach when aggregated to the school-level is that growth statistics will show great sensitivity to changes in the region of the classification cutpoints, but not necessarily to changes in other regions of the score scale (Holland, 2002; Ho, 2008). Unfortunately, the matched longitudinal data needed to illustrate this particular approach in the present context was not available. However, because the CSAP has been vertically scaled, it is at least possible to illustrate in Figure 3 how school-level growth trends can differ when the mean is taken over what is, at the least, an ordinal distribution with multiple levels relative to an ordinal distribution with just two levels.

[4] In fact, there have been some significant demographic shifts in the composition of Creekside students from 2005 to 2011. During the time period, the percentage of Hispanic/Latino students increased from about 27% to 33%. Interestingly, the percentage of students eligible for free and reduced lunch services has actually decreased from 48% to 40%. The former shift might work to excuse the flat trends in Figure 1, but the latter does not.

Figure 3. *Differences in Creekside Trends in Math when the School-Level Outcome is a Scale Score vs. Percent of Student Proficient and Advanced*

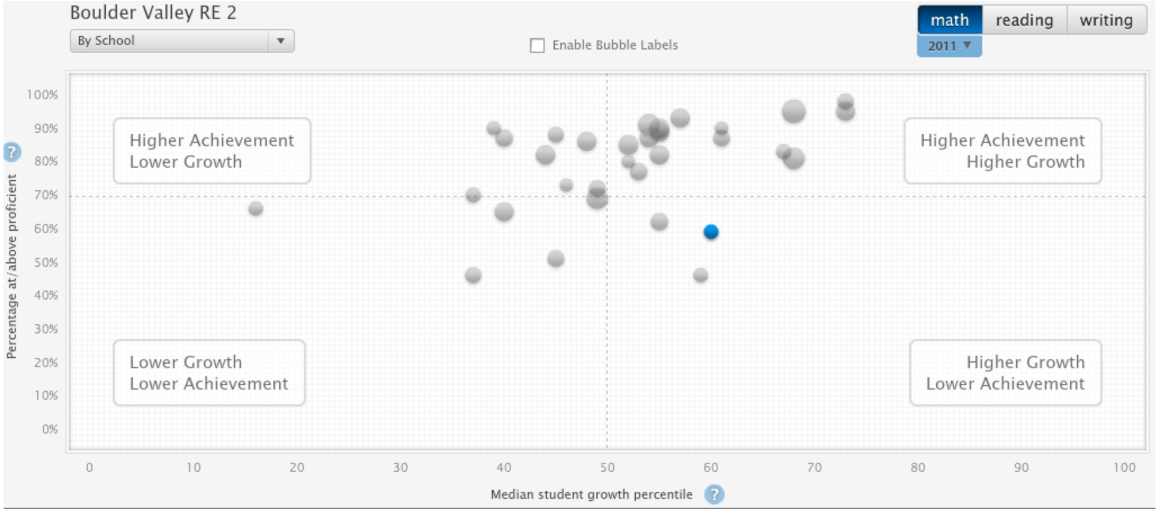


Note: Left Axis = Vertical Scale; Right Axis = % of students proficient or advanced

An approach to growth that preserves more of the information about changes in student performance at all locations along the test score distribution without requiring a vertical scale can be seen in the student growth percentile model (SGPM; Betebenner, 2009) that was first implemented in Colorado and has since been adopted in states such as Massachusetts, Indiana, Wisconsin and Hawaii. In short, the SGPM computes for each student, a conditional test score percentile. This conditional score percentile is found by, in essence, comparing the test score performance of a student in a current grade (e.g., grade X) to all students in the state with the same test score history in all prior grades (e.g., X-1, X-2, etc.) where a standardized test was administered. A student that scores at the 50th percentile of this conditional distribution is one that is inferred to have shown "growth" that represents "one year of learning." An estimate of classroom or school-level growth can then be computed by taking the median over all student growth percentiles for students with test scores in at least two adjacent grades. Note that in this approach the concept of growth is an inference—we infer that if a student has performance that is higher than expected relative to similar students, the reason for this is that they have learned more (i.e., shown more growth) than these similar students[5].

[5] In some cases, when a vertical scale underlies the testing system, this inference can be checked by looking to see whether, in a more absolute sense, such students also have positive score gains from one grade to the

Figure 4. *Plots of School-Level Achievement and Growth, Boulder Valley School District, 2011*



The scatterplot in Figure 4 represents an attempt to combine information about two different dimensions of school "quality": achievement status and growth. Each "bubble" represents a unique elementary school in the Boulder Valley School District and the blue bubble represents Creekside Elementary. The size of each bubble is proportional to the number of students attending a given school. The vertical axis represent the percentage of students in a given school that have been classified as proficient or advanced in math. In contrast, the horizontal axis represents a school's median student growth percentile. In this example, the vertical line at 70 represents the demarcation between schools that are performing better or worse than would be expected given the prior test performance of their students. The horizontal line at 50 represents the threshold set by Colorado for a school to be considered "high achieving." Figure 4 makes it possible to distinguish among four "types" of schools:

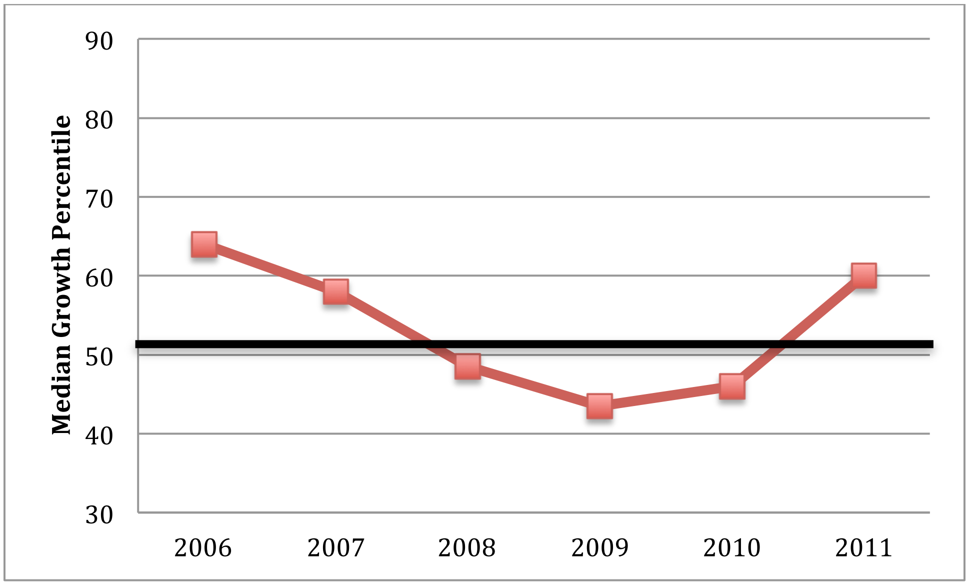
1. Quadrant I: Higher Achievement, Higher Growth
2. Quadrant II: Higher Achievement, Lower Growth
3. Quadrant III: Lower Achievement, Lower Growth

next. Of course, this presumes that the vertical scale is a valid external criterion for growth, and there are often good reasons to be skeptical of such claims (Briggs, 2011).

4. Quadrant IV: Lower Achievement, Higher Growth

The frames of reference for the descriptors "higher" and "lower" are the vertical and horizontal thresholds that have been established by the Colorado Department of Education. The further a school departs from these thresholds, the clearer the designation of a school within each quadrant. Relative to other Boulder Valley School District elementary schools, it can be seen that as of 2011 Creekside lands in Quadrant IV: Lower Achievement, Higher Growth.

Figure 5. *Creekside Trends in Annual Math Growth*



Finally, Figure 5 presents the trends in the median growth percentiles of Creekside students in mathematics from 2006 to 2011. The interpretation from this analysis diverges sharply at points from that based on the pseudolongitudinal data shown in Figures 1 and 2. For example, while trends in math performance showed a steady decline between 2005 and 2011 in Figure 1, this is only mirrored between 2006 and 2009 in Figure 5; from 2009 to 2011 there is evidence of a turnaround.

Using Growth Models to Project the Status of Students in the Future

In Figure 4, information about student growth aggregated to the school level is provided as a visual complement to information about achievement status. It would also be possible to essentially combine the two sources of information by using the growth model to project future trends in achievement status. That is, policymakers may establish fixed expectations for what all students are expected to achieve, as has been the case under No Child Left Behind. When this is the case, the main purpose of the growth model is to give credit to schools whose students are below the progress cutpoints established for any specific grade, but show signs of making rapid growth such that in the future they will meet the expectations that have been established for them. By the same logic, a growth model used for these purposes should indicate the converse—students whose performance is currently adequate but who are losing ground such that they may not meet achievement expectations in the future. This was the impetus for the growth model pilot project that was initiated by Secretary Spellings in 2005 (Spellings, 2005). For many of the states that participated in this project, growth models essentially provided another way to make "safe harbor" if targets based on achievement status were not met.

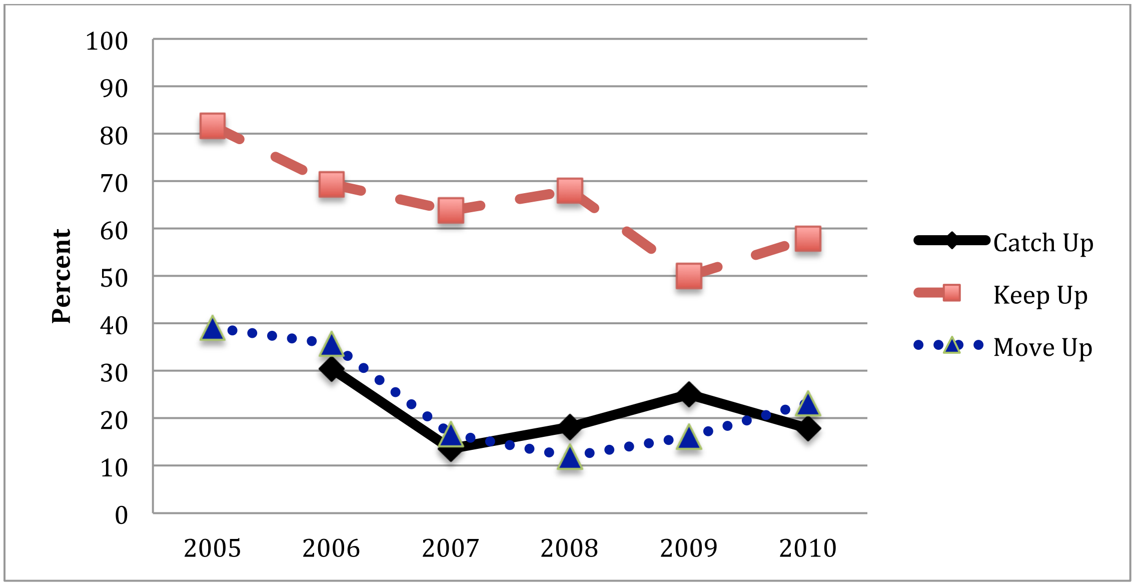
In the context of the SGPM, the key move is to compare, for each student, his or her conditional growth percentile against what is known as an *adequate growth percentile*. An adequate growth percentile represents the conditional growth percentile that needed to have been observed in order for a student to have been classified as proficient or advanced either in the current year or within the next three years[6]. This is computed by using the achievement trends from previous panels of students to predict what is most plausible for the current cohort of students. So for example, imagine that in the past (e.g., 2007) we observed a subsample of grade 5 students who had scored .25 SDs below the score scale threshold for proficiency on the CSAP math test. Three years later, we examine how many of these same students scored above the proficiency threshold. For those that do, we can observe the pattern of conditional growth percentiles that were

[6] For eighth or ninth grade students this is truncated to one or two years because the final grade at which they will be tested is 10th grade.

associated with this shift. We then use this information about the past to estimate the minimum conditional growth percentile trend that a student in the current year (e.g., 2011) would minimally need to pass the proficiency threshold within three years. This represents a student's adequate growth percentile. When a student's adequate growth percentile is well above his or her observed conditional growth percentile, it becomes indicative of the extent the student will need to "catch up" to meet future performance expectations, and the school does not get credit for the student being "on track" to proficiency. However, if his or her conditional growth percentile is higher than the adequate growth percentile, the school does receive credit.

Returning to the Creekside example one last time, Figure 6 plots trends in the projected growth percentiles of grade 4 and 5 students each year who score below, score at, or score above the proficient level on the CSAP math test. There are three different lines of interest. The solid line represents the percentages of grade 4 and 5 students in any given year that score below proficiency, but would be classified as proficient within the next three years if they were to maintain their current conditional growth percentile. These are known as Creekside's "catch up" percentages. The dashed line represents the percentages of grade 4 and 5 students in any given year that score at or above proficiency who would not be classified as proficient within three years if they were to maintain their current conditional growth percentile. These are known as Creekside's "keep up" percentages. Finally, the dotted line represents the percentages of grade 4 and 5 students in any given year that score at proficiency who could be predicted to shift to the advanced category if they were to maintain their current conditional growth percentile. These are known as Creekside's "move up" percentages. It can be seen that in any given year relatively small percentages of Creekside students have had conditional growth percentiles that would enable them to "catch up" or "move up" relative to the state standards that have been set for proficient and advanced performance in math. Indeed, for a significant proportion of students classified as proficient, the growth they are demonstrating would not be enough for them to "keep up[7]", and this is a trend that has worsened over time.

Figure 6. *Growth Projection Approach for Creekside Math Performance*



Projection approaches to growth that are similar to this in spirit, if not in technical details, can be found in the Race to the Top proposals that were funded for Tennessee, Florida and Delaware. The obvious drawback to this approach is that it assumes that the thresholds that have been set for proficiency have been vertically articulated, and that the thresholds are plausible for all students to obtain (no matter what their level of performance in the early grades of testing).

A Note on Normative vs. Absolute Growth Interpretations

Neither the USED or NRC definitions of growth presented earlier specify whether inferences are to be made as a function of absolute or normative changes in student achievement. An "absolute" growth model can be used to answer the question "How much has student achievement changed from one grade to the next?" or "At what rate is

[7] To some extent this is a reflection of the fact that Colorado's standards for proficiency in math actually become harder as students enter middle school. But this would not explain the downward trend over time.

student achievement changing across multiple grades?" A well-known requirement of an absolute growth model is that test scores have been placed onto a vertical scale to adjust for differences in difficulty such that scores across grades can be directly and meaningfully compared. A deeper, and potentially more problematic assumption is that such scales have equal-interval properties (Ballou, 2009; Briggs, 2010, 2011). In contrast, "normative" growth models such as the SGPM do not require a vertically linked scale, only prior test scores that are strongly associated with subsequent test scores. Fundamentally, these models answer the question: "Compared to students with the same prior achievement, is current achievement higher or lower than would be expected?" Normative growth is generally not what a layperson conceptualizes when they are told that a model has been developed to measure growth in student achievement, and this can lead to some confusion when results are being communicated to the public. However, as the illustration above has demonstrated, at the school-level there are some innovative approaches (e.g., adequate growth percentiles) that can be implemented to help assess how much students on the whole have learned and whether the rate of learning appears to be increasing or decreasing. However, these approaches depend greatly on the quality of the horizontal equating and grade to grade standard-setting procedures that underlie the large-scale assessment.

Recap

The trends in achievement status at Creekside paint a worrisome picture that would seem to explain why the school has been sanctioned under the auspices of No Child Left Behind. Yet when we shift from an eye toward trends in status for pseudo-longitudinal cohorts to inferences about growth for longitudinal cohorts, the picture changes considerably. It now appears that the past two cohorts of grade 4 and 5 students at Creekside have test performance that is indicative of significant growth when compared to other students across the state with the same test score histories. On the other hand, relative to the standards the state of Colorado has established for proficient or advanced performance in math, a good case can be made that Creekside students are not growing fast enough. As presented here, a growth model provides an assessment of school-level performance at Creekside that appears more equitable than that which would be given through comparisons of cohort to cohort changes in achievement status. This is because it is based on longitudinal data and takes into account test performance in the past before passing judgment on test performance in the present. It does not, however, simplify the process of reaching a conclusion about the quality of services students at Creekside are receiving. Indeed, the illustration here has been greatly simplified in that I have only focused on math performance and no comparisons as a function of demographic subgroups were made. To come to a summary conclusion about the educational needs at Creekside would surely require further detective work, and in my view, this is how it should be. (For more on this see Briggs, in press).

Making Causal Inferences about Teachers or Schools: "Value-Added Modeling"

There is considerable confusion over the distinction between growth models and value-added models, and the two terms are often used synonymously in discussions of educational accountability. In the National Research Council report *Getting Value out of Value-Added*, Braun et al. (2010) define value-added models (VAMs) as "a variety of sophisticated statistical techniques that use one or more years of prior student test scores, as well as other data, to adjust for preexisting differences among students when calculating contributions to student test performance. (Braun et al. 2010, 1)" According to Harris (2009), "the term is used to describe analyses using longitudinal student-level test score data to study the educational input-output relationship, including especially the effects of individual teachers (and schools) on student achievement." Given these definitions, the clearest distinction between a growth model and a value-added model is inferential intent. A growth model is first and foremost descriptive, and makes no explicit attempt to "isolate" the contribution of teachers or schools to student achievement. By contrast, it can be argued that a value-added model is specified first and foremost with the intent of estimating the causal effects of teacher (or schools) on students[8]. This distinction between growth and value-added is easy to blur because the

[8] For example, SAS makes the following explicit claim in its marketing of its Educational Value Added Assessment System (EVAAS): "It is much more than teacher or classroom level analyses; it assesses the

moment that student-level growth statistics are aggregated to the classroom or school levels, it will often be unavoidable that high or low values are attributed to teachers or schools in a causal manner. However, while all growth models applied to students in educational settings may well encourage value-added inferences, in the absence of additional evidence, there is no compelling reason a priori to believe that these inferences—in the absence of additional detective work—will be valid.

At this point there have been a number of excellent reviews written on issues surrounding the specification and use of value-added models (Braun et al., 2010, McCaffrey et al, 2003; McCaffrey, Han & Lockwood, 2009; OECD, 2008; Harris, 2009). I briefly focus attention on three distinct approaches that might be taken to use PARCC test scores to make value-added inferences. In a departure from the Creekside example above, I illustrate each model for the case in which classrooms rather than schools are the units of analysis.

The Student Growth Percentile Attribution Approach

The SGPM (Betebenner, 2009) quantifies student achievement relative to prior achievement in the metric of cumulative score percentiles. A student growth percentile (SGP) is computed for each student in a given grade with at least two years of consecutive tests scores using quantile regression (Koenker & Hallock, 2001). No imputation is performed to include students with missing values. Quantile regression can be conceptualized as an elaboration of the widely known and applied Ordinary Least Squares linear regression from a case in which one is interested in parameterizing trends in conditional means, to one in which one is interested in parameterizing trends in many different conditional quantiles. An advantage of the quantile regression underpinnings of the SGPM is that the approach does not assume linearity in its regression functions, is insensitive to outliers, the presence or absence of vertical links between the score scales

effectiveness of districts, schools and teachers, as well as provides individual student projections to future performance. SAS EVAAS for K-12 provides precise, reliable and unbiased results that other simplistic models found in the market today cannot provide." Retrieved August 25, 2011 from http://www.sas.com/govedu/edu/k12/evaas/index.html.

from grade to grade, and whether or not the score scale for any given grade is interval or ordinal. Specialized software is required to specify and estimate a quantile regression, but such software is publicly available and the SGP package is free through the R statistical environment.

The SGPM is a perfect example of why the terms growth model and value-added model are sometimes used interchangeably: because student-growth percentiles can be easily aggregated to the classroom and school-levels, they are often given a de facto attribution as value-added. Yet the model was not developed as a competitor to other value-added models, and the statistics it produces at the classroom or school level were never meant to represent a direct estimate of the causal effect of teachers or schools on student achievement[9]. So while it may well be used as though it were a value-added model, and it may even lead to many of the same rankings and conclusions as some value-added models, this is not its purpose, which is to provide a descriptive characterization of conditional student achievement that supports inferences about growth. Nonetheless, it seems important to note that just because value-added models are *intended* as tools to isolate the causal effects of teachers and schools on student achievement this does not necessarily mean they are any more capable of accomplishing this relative to the SGPM (Betebenner, Wennig & Briggs, 2011). In fact, in two instances where the teacher and school rankings based on the SGPM and a widely used VAM (i.e., the EVAAS, see below) have been compared, the typical correlations have been greater than 0.8 (Briggs & Betebenner, 2009; Wright, 2010).

The Production Function Approach

Let Y_{ig} represent the composite end of year test score on a PARCC assessment for student (i) in grade (g), in a classroom with teacher (t) and school (s). The VAM is specified as

$$Y_{ig} = \alpha_g + \beta' X_{ig} + \gamma' Z_{ig}^{(ts)} + \sum_t \theta_t D_{ig} + \varepsilon_{ig} . \tag{1}$$

[9] For more on this issue, which is really quite subtle, please see Briggs (in press).

The covariates in this model are captured by X_i which represents PARCC test scores from prior grades[10] , and $Z_{ig}^{(ts)}$, which could represent any number of student, classroom or school-specific variables thought to be associated with both student achievement and classroom assignment. The key parameter of interest is the fixed effect θ_t, which represents the effect of the current year's teacher on student achievement (i.e., the model above includes a dummy variable indicator for each teacher in the dataset).

The validity of the model hinges upon the assumed relationship between the unobserved error term ε_{igts} and D_{ig}. If, conditional upon *X* and *Z*, ε_{igts} and D_{ig} are independent, in theory it is possible to obtain an unbiased estimate of θ_t. In other words, if one can control for the variables that govern the selection process whereby higher or lower achieving students land in certain kinds of classrooms, then one can approximate the result that would be obtained if students and teachers had been randomly assigned to one another from the outset. This is controversial proposition, and much of the debate over the use of VAM for teacher accountability has focused on (a) the nature and number of covariates that need to be included in *X* and *Z*, and (b) evaluating the extent to which adding more variables or student cohorts serves to reduce bias in θ_t.

The production function model has a long history in the economics literature (Hanushek & Rivkin, 2010; Todd & Wolpin, 2003), and helps to explain why this has been the preferred specification approach among economists who have contributed the VAM literature. This is the specification that underlies the VAM used by Wisconsin's Value-Added Research Center, which has taken an active role marketing its services to urban school districts across the country (e.g. New York City, Milwaukee, Los Angeles). There is ongoing debate over whether certain VAMs from this tradition can be used to support unbiased causal inferences about teacher effects. For optimistic perspectives, see Kane & Staiger, 2008; Koedel & Betts, 2009; and Goldhaber & Hansen, 2010. For more pessimistic assessments, see Ballou, 2009; Rubin, Stuart & Zanutto, 2004; and Rothstein, 2009; 2010.

[10] Depending on the current year grade of the student, the number of available prior test scores in the same subject could range anywhere from 1 (if current year grade of student i is 4), to 8 (if current year grade of student i is 12).

The EVAAS Approach

The Educational Value-Added Assessment System (EVAAS; Sanders, Saxton & Horn, 1997) is a multivariate longitudinal mixed effects model. While a detailed presentation is outside the scope of this paper, a key point of differentiation between it and the approaches presented above can be seen by writing out the equation for a single test subject in parallel to equation 1

$$Y_{ig} = \alpha_g + \sum_{g^* \leq g} \theta_{g^*} + \varepsilon_{ig}. \qquad (2)$$

In contrast to the fixed effects specification above, under this approach teacher effects for a given grade are cast as random variables with a multivariate normal distribution such that $\theta_{g^*} \sim N(\mathbf{0}, \tau)$. Only the main diagonal of the covariance matrix is estimated (i.e., teacher effects are assumed to be independent across grades). The student-level error term is also cast as a draw from a multivariate normal distribution with a mean of 0, but the covariance matrix is left unstructured. The EVAAS is often referred to as the "layered model" because a student's current grade achievement is expressed as a cumulative function of the current and previous year teachers to which a student have been exposed. For example, applying the model above to longitudinal data that span grades 3 through 5 results in the following system of equations:

$$\begin{aligned} Y_{i3} &= \alpha_3 + \theta_3 + \varepsilon_{i3} \\ Y_{i4} &= \alpha_4 + \theta_3 + \theta_4 + \varepsilon_{i4} \\ Y_{i5} &= \alpha_4 + \theta_3 + \theta_4 + \theta_5 + \varepsilon_{i5} \end{aligned}$$

In the model above no teacher effects can be computed for grade 3 because they are confounded with variability in student achievement backgrounds. In contrast, when certain assumptions hold it is possible to get an unconfounded effect for the grade 4 teacher. This can be seen by substituting the first equation into the second equation in the system such that $Y_{i4} - Y_{i3} = \alpha_4 - \alpha_3 + \theta_4 + \varepsilon_{i4} - \varepsilon_{i3}$. This shows that the sufficient statistic for estimates of teacher effects under the EVAAS are test score gains from one grade to the next. It is for this reason that the EVAAS (and other mixed effect modeling approaches related to it) has long been presumed to require test scores that had been vertically scaled (Ballou, Sanders & Wright, 2004; McCaffrey et al., 2003).[11]

This simplified presentation may obscure two significant aspects of the EVAAS that contribute to its purported ability to reliably and validly "disentangle" the influence of teachers from other sources of that influence student achievement. In particular, the EVAAS

- makes use of panel data for up to five years of test data per student and three student cohorts per teacher; and
- models multiple test subject outcomes jointly as a multivariate outcome.

The EVAAS has the longest history as an approach being used for the purposes of educational accountability. Though it has been criticized because it does not control for additional covariates beyond a student's test score history (Kupermintz, 2003), teacher value-added estimated by the EVAAS with and without student-level covariates have been shown to be strongly correlated (Ballou, Sanders & Wright, 2004). A more equivocal issue is whether or not one should control for classroom or school-level characteristics (McCaffrey et al, 2004). Controlling for classroom characteristics can lead to overadjustments of teacher effect estimates; controlling for school fixed effects will restrict teacher comparisons to a within-school reference population. Finally, note that the EVAAS assumes that the effects of students' teachers in the past persist undiminished into the future. McCaffrey et al (2004) and Lockwood, McCaffrey, Mariano & Setodji (2007) have demonstrated empirically that this may not be a viable assumption in the context of teachers; Briggs & Weeks (2011) show that it is probably not viable in the context of schools. However, this issue only has an impact on the precision of value-added estimates, not with accuracy.

[11] As it turns out, while it does matter how a test has been scaled (Briggs & Weeks, 2009), it will generally make little difference to value-added rankings of teachers or schools whether the tests have or have not been vertically linked. I demonstrate this in a subsequent section.

Assessment Design Factors that Could Impact Inferences about Growth and/or Value-Added

Vertical Scaling[12]

A vertical scale is not a prerequisite when tests are being used to make value-added inferences. This is true not only in the case of the SGPM and production function approaches where it may seem self-evident, but is also generally true for approaches (like the EVAAS) that depend upon gain scores or repeated measures as sufficient statistics. From a purely empirical standpoint, the presence or absence of a vertical scale when implementing a reduced form version of the EVAAS has been shown to make almost no difference to the normative rankings of teacher or schools from a value-added model (Briggs & Weeks, 2009; Briggs & Betebenner, 2009; Briggs & Domingue, in progress). But a theoretical explanation for this can be established as well, and it seems worthwhile to do so here.

Imagine a pair of vertically scaled assessments taken over the course of two years. Denote the pair of scale scores (that have been scaled via Item Response Theory but not yet placed on the vertical scale) by *y* and *x*, where *x* comes from grade *g* in year *t* and *y* comes from grade *g* + 1 in year *t* + 1. Vertical scaling imposes a grade-specific linear transformation on *x* and *y* where the linking constants are usually estimated iteratively using the Stocking-Lord Algorithm (Stocking & Lord, 1983). After the two tests have been vertically scaled we have

$$x' = \alpha_0 + \alpha_1 x$$
$$y' = \beta_0 + \beta_1 y.$$

It is easy to show that when using some version of the VAM from equation 1, there will be a perfect correlation between estimates of value-added whether one is using *y* and *x* or y' and x' in the model. To keep the proof simple without any loss of generality, we can rewrite equation 1 for the case where we are estimating the value-added of a school on its students in grade 5 with only a single prior year test score in grade 4 as

[12] Part of this section draws from a manuscript in progress by Briggs & Domingue. The mathematical argument being presented was first established by Ben Domingue.

$$y_i = \gamma_0 + \gamma_1 x_i + \sum_t \theta_t D_i + \varepsilon_i \,. \tag{3}$$

Now consider the same model after the two scales have been vertically linked.

$$\beta_0 + \beta_1 y_i = \gamma_0 + \gamma_1(\alpha_0 + \alpha_1 x_i) + \sum_t \theta_t D_i + \varepsilon_i \tag{4}$$

With a little algebra (4) can be rewritten as follows

$$y_i = \left(\frac{\gamma_0 - \beta_0}{\beta_1}\right) + \left(\frac{\gamma_1(\alpha_0 + \alpha_1 x_i)}{\beta_1}\right) + \frac{\sum_t \theta_t D_i}{\beta_1} + \frac{\varepsilon_i}{\beta_1}.$$

It follows that the value-added parameters θ_t and $\frac{\theta_t}{\beta_1}$ from (3) and (4) have a perfect linear relationship. Now consider the same proof in the case where the outcome variable of interest is a gain score. Before vertical links have been established, we have

$$y_i - x_i = \gamma_0 + \sum_t \theta_t D_i + \varepsilon_i . \tag{5}$$

Once again, consider the same model after the two scales have been vertically linked:

$$\beta_1 y_i - \alpha_1 x_i = \gamma_0 - \alpha_0 - \beta_0 + \sum_t \theta_t D_i + \varepsilon_i . \tag{6}$$

In this case, unless the two linking constants α_1 and β_1 that affect the variability of the scales are identical, there is no guarantee that the value-added estimates from (5) will be linearly related with those from (6). However, using grade 3 through 8 data from Colorado we have examined the correlation between school-level estimates when we fix α_1 at 1 (a common identification constraint imposed on the base grade of a vertical scale) and let β_1 vary. For values of β_1 between .9 and 1.1, our correlations between school-level estimates are 0.97. Only when the values drop (increase) as low (high) as .8 (1.2) do we see a noticeable drop in our correlations to .89. In our empirical work with vertical scales based on the grade span between 3 and 8, we have found that the estimates for these constants from grade to grade range between about .95 and 1.05. Hence in most realistic testing contexts the decision to link score scales vertically is unlikely to have a significant impact on the value-added rankings of teachers or schools.

The discussion above is not meant to imply that vertical scales are not desirable, or that PARCC should not consider establishing some—only that this needs to be done for the right reasons. It is possible that choices in vertical scaling can have a major impact when they are used as a basis for simple parametric linear models that project student achievement into the future. For example, in some states, a vertical scale is used as a means of setting vertically articulated cut-points across grades through the process of standard-setting. Since projections of student achievement are evaluated relative to these cutpoints, if two different vertical scales led to different cutpoint locations, this could change the cumulative distribution of students below a given cutpoint. Beyond this, Kolen (2011) has argued that a vertical scale can be useful as a means for connecting a single number associated with a school, classroom or student back to meaningful statements about the content of the test that has or has not been learned. As noted earlier, only a vertical scale can support statements about student-level growth in terms of magnitudes that are analogous to the sorts of statements that would be made about a child's growth in height. Unfortunately, not all vertical scales are created equally (Briggs, 2009; 2011), and there are good reasons to be wary about the claims they can support.

Nonetheless, there is some reason for optimism about the prospects for vertical scaling with the PARCC assessments. In contrast to most existing large-scale assessments, the PARCC tests will be designed on the basis of a standards framework in which focused attention to growth over time in "highlighted domains" was a guiding principle. If careful thought is given ahead of time to the proper design and validation of vertical scales, they could play an important role in (a) communicating growth trends at the student-level in a manner that is intuitive to stakeholders, and (b) providing complementary information that could serve as a criterion-referenced validity check on normative value-added inferences.

One way this might be accomplished in mathematics would be to attempt to create staggered vertical scales for a small set highlighted domains. For example, see Figure 7. In elementary school, vertical scales could be created within the domains of "numerical operations: fractions" and "measurement". By middle school, growth would be measured using new vertical scales within the domains of "expressions & equations" and "geometry." Because no vertical scale would span more than 3 to 4 grades, it would feasible to implement a scaling design in which the same external form was administered to all students across a given grade span containing the same set of items. In ELA, a single vertical scale for a well-understood construct like reading comprehension would appear to be a more viable possibility (Briggs, 2011; Stenner, Burdick, Sanford, & Burdick, 2006; Stenner & Stone, 2010).

Figure 7. *Staggered Vertical Scales in Mathematics*

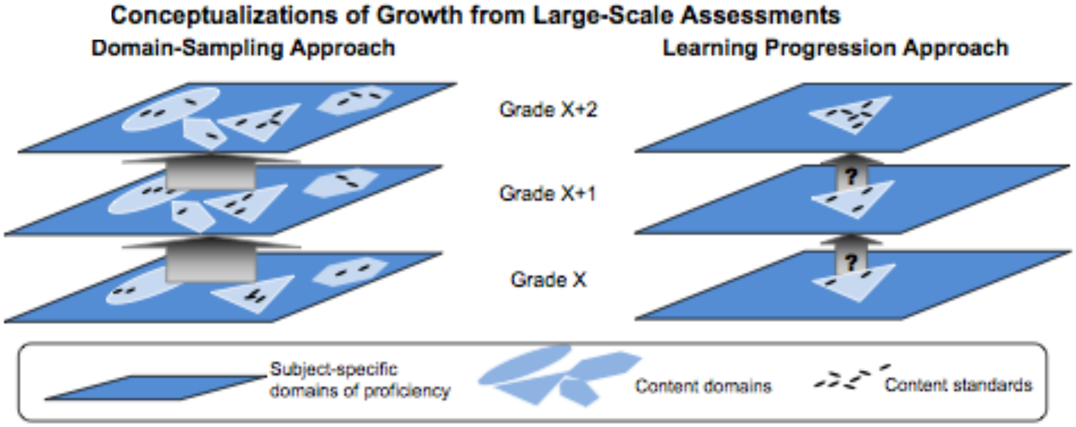
Grade	Fractions	Measurement	Expressions & Equations	Geometry
8			X	X
7			X	X
6	X		X	X
5	X	X		
4	X	X		
3	X	X		

Construct Conceptualization

A key design feature of the PARCC tests will be the construct conceptualizations within the mathematics and ELA subject areas. Figure 8 shows two different growth interpretations associated with two different conceptualizations of the construct of measurement. The left side of Figure 8 contains planes that are intended to demonstrate what it means to be “college and career ready” in a given subject area (e.g., mathematics) at a given grade level (e.g., grade 3). Within each plane are light-colored shapes, and within each shape is a series of dots. The shapes are meant to represent different “content domains” (e.g., Numerical Operations, Measurement & Data, Geometry); the dots represent domain-specific performance standards that delineate grade-level expectations for students (e.g., within the domain of Measurement & Data: “Generate measurement data by measuring lengths using rulers marked with halves and fourths of an inch.”). This sort of taxonomy has traditionally been used to deconstruct the often amorphous notion of “mathematical ability” into the discrete bits of knowledge, skills, and abilities that should, in principle, be teachable within a grade-level curriculum. Such an approach facilitates the design of grade-specific assessments because test items can be written to correspond to specific statements about what students should know and be able to do.

The construct of measurement in such designs is not a well-understood psychological attribute of the test-taker, but a composite of many, possibly discrete, KSAs. I refer to the assessment design implied by the left side of Figure 8 as the *domain-sampling* approach.[13]

Figure 8. *Different Construct Conceptualizations and Implications for Growth*


Conceptualizations of Growth from Large-Scale Assessments
Domain-Sampling Approach
Learning Progression Approach
Grade X+2
Grade X+1
Grade X
?
?
Subject-specific domains of proficiency
Content domains
Content standards

Under the domain-sampling approach, the intent is for growth to be interpreted as the extent to which a student has demonstrated increased mastery of the different domains that comprise mathematical ability. This is indicated by the single arrow indicating movement from the plane for a lower grade to the plane for a higher grade. Note that if both the domains and the content specifications within each plane change considerably from grade to grade, then it becomes possible for students appear to "grow" even if distinct content is tested across years. In the best case scenario for growth inferences, considerable thought has been put into the vertical articulation of the changes among content domains, clusters, and standards from grade to grade. For example, according to the CCSS, a composite "construct" of mathematical ability could be defined from grade to grade as a function of 5 content domains and 6 skills domains. Yet this leaves ample room for growth in terms of the composite to have an equivocal interpretation depending upon the implicit or explicit weighting of the domains in the assessment design and scoring of test items. Furthermore, the number of items required

[13] For complementary perspective see the recent article by Markus & Borsboom (2011).

to make inferences about *all* CCSS domains at one point in time along with change over time is likely to be prohibitive.

A different basis for construct conceptualization comes from what I refer to as the *learning progression* approach. Learning progressions have been defined as empirically grounded and testable hypotheses about how students' understanding of core concepts within a subject domain grows and become more sophisticated over time with appropriate instruction (Corcoran, Mosher, & Rogat, 2009). The key idea shown in the right panel of Figure 8 is the presence of an implicit *hypothesis* about the nature of growth: the way that students' understanding of some core concept or concepts *within the same domain* is expected to become qualitatively more sophisticated from grade to grade. The notion that this constitutes a hypothesis about growth to be tested empirically is represented by the question marks placed next to the arrows that link one grade to the next. The learning progression approach to growth values the assessment of depth of knowledge within a single domain over the assessment of breadth of knowledge across multiple domains. Hence there is a cost to following this approach exclusively as it may reduce the ability of a testing system to "assess the full range of the CCSS."

To illustrate this idea more concretely, consider a specific example of something that might serve as an initial basis for a learning progression hypothesis: the concept "represent and interpret data" that can be found in the "Measurement and Data" domain in the CCSS. The standards associated with this concept each time it appears across grades 1 through 5 are shown in Figure 9 below.

Figure 9. *Example of a Learning Progression Hypothesis That Could be Made on the Basis of the Measurement & Data Domain of the Common Core Standards*

Grade	Grade Level Performance Expectations for "Represent and Interpret Data"
5	• Make a line plot to display a data set of measurements in fractions of a unit (1/2, 1/4, 1/8). Use operations on fractions for this grade to solve problems involving information presented in line plots. *For example, given different measurements of liquid in identical beakers, find the amount of liquid each beaker would contain if the total amount in all the beakers were redistributed equally.*
4	• Make a line plot to display a data set of measurements in fractions of a unit (1/2, 1/4, 1/8). Solve problems involving addition and subtraction of fractions by using information presented in line plots. *For example, from a line plot find and interpret the difference in length between the longest and shortest specimens in an insect collection.*
3	• Draw a scaled picture graph and a scaled bar graph to represent a data set with several categories. Solve one- and two-step "how many more" and "how many less" problems using information presented in scaled bar graphs. *For example, draw a bar graph in which each square in the bar graph might represent 5 pets.* • Generate measurement data by measuring lengths using rulers marked with halves and fourths of an inch. Show the data by making a line plot, where the horizontal scale is marked off in appropriate units—whole numbers, halves, or quarters.
2	• Generate measurement data by measuring lengths of several objects to the nearest whole unit, or by making repeated measurements of the same object. Show the measurements by making a line plot, where the horizontal scale is marked off in whole-number units. • Draw a picture graph and a bar graph (with single-unit scale) to represent a data set with up to four categories. Solve simple put-together, take-apart, and compare problems using information presented in a bar graph
1	• Organize, represent, and interpret data with up to three categories; ask and answer questions about the total number of data points, how many in each category, and how many more or less are in one category than in another.

In this illustrative learning progression the ability of students to represent and interpret data is expected to increase in sophistication from grade to grade: In grade 1 students are expected to make frequency comparisons across a limited number of nominal categories; by grade 2 they are expected to be able to make relatively crude ordinal measurements of length; by grade 3 they are expected to make these ordinal comparisons more precise; and by grade 4 they are, for all intents and purposes, making continuous measurements of length. Note that when the ability to represent and interpret data is viewed in terms of the continuum above, there are important implications for the design of items that would be used to assess student understanding. Say that grade 3 students are given items that indicate whether or not they can represent and interpret data so as to make measurements to the nearest ¼ of an inch. If a student fails to master the items, does this mean he/she is struggling with the prerequisite skills listed for grade 2 or grade 1? Conversely, if a student demonstrates mastery of the items, this doesn't help us identify whether he/she might have an understanding that is even higher on the hypothesized progression. The upshot of all this is that to make meaningful inferences about growth, students need to be assessed from grade to grade with items that are both below (easier) and above (harder) the level of understanding anticipated on the basis of the standards This is a fundamental way that a learning progression approach to assessment design differs from a domain-sampling approach.

A message that has been delivered time and time again by the PARCC TAC is that the validity of any large-scale assessment will hinge upon the ability of test developers to be clear and concrete about what it is that is being measured. From the perspective of designing assessments that can support inferences about growth, this message must be extended to a demand that test developers be clear and concrete about what is expected to change over time. In this regard, because vertical linkages are already evident in the progression of domains and standards across grades, the CCSS are a step in the right direction as a blueprint for an assessment system focused on growth. The two ways of conceptualizing the construct of measurement I have presented here, domain-sampling and learning progression, are by no means exhaustive or mutually exclusive. One could imagine a large-scale assessment in which items associated with a learning progression are embedded as a stratum within a larger sample across domains. However, I would argue that the learning progression approach is much more likely than the domain sampling approach to lead to a theoretically defensible basis for a vertical score scale.

Measurement Error and Inferences about Growth

Regardless of the modeling approach being taken, growth inferences for students at the low and high ends of the score distribution will be biased downward in the presence of measurement error. It may be easiest to think of this issue in term of floor and ceiling effects. Floor effects will occur when the easiest items on an assessment are still too difficult for students at the low end of the total score distribution. Ceiling effects will occur when the hardest items are still too easy for students at the high end of the total score distribution. In item response theory, it is well understood that measurement error at a particular location of a test score scale has an inverse relationship to the number of items at the same location. Because of constraints in testing time and a recognition that the bulk of test-takers will not be located at the extremes of the score distribution, most large-scale assessments have an information function with the typical inverted U shape. If the PARCC tests were to conform to this, then making student-specific inferences about growth from year to year will remain problematic for students at the low and high ends of the distribution.

The best way to avoid floor and ceiling effects is to increase the possibility of "out of level" testing. The idea would be to test students where they are, not where they should be. Though PARCC is not planning a computer adaptive testing structure, it might still be possible to "route" students to more targeted test forms for their end-of-year testing on the basis of their performance on one of the earlier through-course tests. It does not seem advisable however, to hold the PARCC tests to the criterion of the same (low) SEM for students no matter where they happen to be located in the score distribution as a computer adaptive test. One might argue that for the extreme cases in which *all* the students in a teacher's classroom are at least two SDs below the mean of the test score distribution, there are much bigger issues than measurement error to worry about. My (admittedly limited) empirical work on this issue suggests that such cases will be relatively rare. Using the data from the Los Angeles Unified School District (LAUSD; Briggs & Domingue, 2011) we can ask how many grade 3-5 teachers have classrooms in which no student has a test score that is higher than 2 SDs below average. There were between 1,694 and 1,861 distinct classrooms in the LAUSD in 2009. Out of this number, there is a not a single example of a classroom that would meet the 2 SD criterion. If we loosen the criterion to a classroom in which no student has a test score higher than 1 SD below average, we find that for reading outcomes there is one grade 3 classroom, two grade 4 classrooms and one grade 5 classrooms where this applies. For math, there is a single grade 4 and 5 classroom. Needless to say, as a percent of total, this is less than .05% of all LAUSD classrooms. So while there is considerable variability in test score performance by classroom in Los Angeles elementary schools (the proportion of total variance that is between classrooms rather than within them are between about .35 and .36 in math and .38 and .40 in reading), in this particular example it seems unlikely that student tracking by ability is so extreme that measurement error would become a primary technical objection to student or classroom-level inferences about growth.

Measurement Error and Value-Added Indicators

Measurement error can pose problems for stability of value-added indicators in two different ways at two different levels. First, to the extent that the student-level regression equations at the foundation of a value-added analysis includes prior year test scores as "control" variables, measurement error in these observed scores will lead to an attenuation of *all* regression coefficients in the model (Fuller, 1987; Buonocarsi, 2010). Second, regardless of the quality of instruction to which they are exposed, it may be the case that some cohorts of students are simply "better" or "worse" than others[14]. Given this, when teachers or schools are the units of analysis to which inferences are being made, it has been argued that some portion of the observed variability in estimates of value-added can be explained by chance (Kane & Staiger, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009). The key distinction here is that measurement error at the student level is assumed to have a functional relationship with the number of test items that students have been administered; at the teacher or school level, measurement error is assumed to have a functional relationship with the number of students. The analogy here is essentially that students are to schools what items are to students. Taken together, both of these sources of measurement error could explain the phenomenon in which value-added estimates appear to "bounce" up and down in a volatile manner from year to year—even if the true value were actually constant over time.

[14] This is sometimes described as "sampling error" rather than measurement error, but the concept is the same.

Measurement Error at the Classroom or School Levels

The year to year correlation of teacher value-added has been found to be weak to moderate, ranging from about 0.2 to 0.6 (Goldhaber & Hansen, 2008; McCaffrey et al., 2009). Kane & Staiger (2002; 2008) have argued that such intertemporal correlations can be interpreted as an estimate of reliability, in which case any intertemporal correlation less than 0.5 would imply that more than half of the variability in value-added can be explained by chance unrelated to characteristics of teacher or school quality that persist over time. After conducting a simulation study, Schochet & Chiang (2010) conclude that 35% of teachers are likely to be misclassified as either effective or ineffective when classifications are based on a single year of data.

Three adjustments are typically made, sometimes in tandem, to account for the instability of value-added indicators. One adjustment is to increase the number of years of data over which value-added is being computed. Schochet & Chiang (2010) find that going from one to three years of data reduces the error rate for teacher effectiveness classifications in their simulation from 35 to 25%. Using empirical data from Florida, McCaffrey et al. (2009) find that going from one to three years of data increases, on average, the reliability of value-added estimates for elementary school and middle school teachers from 0.45 to 0.55 and 0.56 to 0.66 respectively. Two related adjustments are to use these estimates of reliability to "shrink" value-added estimates back to the grand mean (i.e., the average value-added of all teachers in the system), and/or to compute a confidence interval around each value-added estimate.

When the reliability of value-added is low, it will typically be a mistake to attempt to classify teachers or schools into more than three categories (e.g., significantly below average, average, significantly above average). In such instances if teachers are instead classified into quintiles of the value-added distribution (five equally spaced categories instead of three unequally spaced categories), misclassification rates are likely to increase dramatically (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Briggs & Domingue, 2011).

Measurement Error at the Student Level

Considerably less research has been done to evaluate the impact of student-level measurement error on value-added analyses. The problem only appears to be relevant to regression models in which prior year test scores are included as independent variables. Such cases lead to the classic "errors in variables" problem that is well-understood in the econometrics literature. Though there are many possible adjustments that could be used as a correction to this problem (c.f., Fuller, 1987; Buonacorsi, 2010), the adjustments that have been applied in the literature to date (c.f., Buddin & Zamarro, 2009; Rothstein, 2009) have been based on the assumption of constant measurement error across the test score distribution, an assumption that is clearly unrealistic given the way that large-scale assessments are designed (see next section). While it is clear that the failure to adjust for the error in variables problem can have a significant impact on value-added inferences, the practical impact of imposing linear instead of nonlinear adjustments is unclear. This is likely to be an active area for research studies in the coming years.

One subtle issue that might have an impact on VAM usage comes from a footnote in Brennan's recent White Paper "Using Generalizability Theory to Address Reliability Issues for PARCC Assessment." Brennan (2011) writes in footnote 7, "To put it bluntly, coefficient α is not likely ever to be a defensible statistic for characterizing the reliability of scores for a PARCC assessment." In the production function VAM specifications, it is typical for econometricians to make adjustments that take into account measurement error in test scores included as covariates on the right-hand side of the equation. Most of these adjustments involve some use of Cronbach's alpha to estimate a single standard error of measurement. To the extent that Brennan's assertion is correct, this may well have an impact on the adjustments made for measurement error in the production function VAM approach, especially if there are significant differences in conditional standard errors of measurement.

Horizontal Equating

Growth models that are used to supplement the evaluation of achievement status (e.g., "college and career readiness by the end of high school") through projections of student achievement into the future depend on the assumption that test scale scores for the same grade are, at a minimum, horizontally comparable across time. At present there is considerable empirical evidence to suggest that some large-scale assessments have had problems with their horizontal equating. On one state's test my students and I have found examples of shifts in horizontal "ability" across years that were larger than those found vertically across grades. Meyer & Dokumaci (2009) expressed similar concerns in the context of horizontal shifts found for Wisconsin's state test. While substantial shifts upward might not be surprising in the early years following the implementation of a new assessment, in some states these shifts appear to continue in subsequent years, in both directions. For more on this issue see the PARCC White Paper by Luecht & Camara (2011).

Testing Window and Days of Instruction

One design issue that has only recently come to my attention is the timing and length of the testing window during which students in participating states will be expected to have taken the PARCC assessments. Assume for the moment that the PARCC tests will in fact be sensitive to instruction in the sense that ceteris paribus, a student exposed to an additional five days (or 10 days, or 15 days, etc.) of focused instruction that is aligned to the CCSS (i.e., a unit of adding/subtracting fractions with different denominators) will score higher, on average, then a student who has not. It follows that when test scores are being used to evaluate teachers and schools it is not likely to be a matter of indifference whether a given classroom or school is comprised of students that were tested at the beginning or end of the testing window. The key variable is not so much the testing window, but the number of days of instruction to which students have been exposed before they are tested. If this varies considerably by classrooms, school districts and/or states, then to the extent that value-added comparisons are to be made across these different units of analysis, it will be necessary to include this as a covariate in a VAM (e.g., production function approach) or as a inclusion criterion for the teachers and schools eligible to be compared with respect to value-added (e.g., EVAAS approach). In general, then having a shorter time window will mitigate this concern. The tradeoff, of course, is that having a shorter time window may limit the PARCC's ability to implement a large-scale assessment with truly innovative features, as accommodating such features (such as performance tasks that span multiple days) may require a longer testing window.

Recommendations

1. The first priority for PARCC needs to be to design large-scale assessments that allow for student-level inferences about what students know, can do, and have learned in the subject areas of mathematics and English Language Arts. And these inferences should form the basis for judgments about college and career readiness. Inferences about classroom or school-level growth and value-added should be (distant) secondary priority because these are generally not something that can be directly influenced by assessment design decisions.
2. Inferences about growth will be most interpretable when the constructs of measurement have properly conceptualized. To this end it is important to be explicit about the KSAs that are and are not changing from grade to grade in the content areas of math, reading and writing. This notion was illustrated through the learning progression approach to assessment design shown in Figure 9.
3. In support of this, PARCC should find a way to embrace a considerable degree of "out of level" testing. Items will need to be administered to certain students in a particular grade that may well be too easy to too hard for them in a traditional sense. This may mean that some items that would be rejected under a traditional review of classical item statistics would be maintained. An increase in out of level testing will have the effect of decreasing floor and ceiling effects of grade-specific tests.
4. Conditional on #2 and #3 above, PARCC should develop vertical scales. In mathematics, these vertical scales should be targeted to specific highlighted domains across a limited span of grades. In reading, it might be feasible to develop a vertical scale for the construct of reading comprehension that spans grades 3 through 12. There does not seem to be a sufficient basis for the

development of a vertical scale in writing. I want to emphasize the point that a vertical scale should only be developed given the constraints I have presented here, and if clear criteria are established for how the quality of the scale will be evaluated.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95-135.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissitz (Ed) *Value added models in education: Theory and applications*, 272–303.

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

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

Betebenner, D., Wennig, R. & Briggs, D. C. (2011). Student growth percentiles and show leather. Education News Colorado. http://www.ednewscolorado.org/2011/09/13/24400-student-growth-percentiles-and-shoe-leather

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. *Princeton, NJ: Educational Testing Service. Retrieved February*, *27*, 2008.

Braun, H, Chudowsky, N, & Koenig, J (eds). (2010) *Getting value out of value-added. Report of a Workshop*. Washington, DC: National Research Council, National Academies Press.

Briggs, D. C. & Betebenner, D. (2009) Is Growth in Student Achievement Scale Dependent? Paper presented at the invited symposium "Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues" at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

Briggs, D. C. & Weeks, J. P. (2009) The sensitivity of value-added modeling to the creation of a vertical scale. *Education Finance & Policy*, 4(4), 384-414

Briggs, D. C. (2010). The problem with vertical scales. Paper presented at the 2010 Annual Meeting of the American Educational Research Association, Denver, CO, May 3, 2010.

Briggs, D. C. (2011). Measuring growth with vertical scales. Working Paper. In review at *Journal of Educational Measurement*.

Briggs, D. C. (in press). Making value-added inferences from large-scale assessments. In Simon, M., Ercikan, K., & Rousseau, M (Eds.) Improving Large-Scale Assessment in Education: Theory, Issues and Practice. London: Routledge.

Briggs, D. C. & Domingue, B. D. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by the Los Angeles Times. National Education Policy Center. http://nepc.colorado.edu/publication/due-diligence.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*.

Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods, and Applications. New York: Chapman and Hall/CRC.

Fuller, W. A. (1987). *Measurement Error Models*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Goldhaber, D. & Hansen, M. (2010). Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions. CALDER Working Paper 31.

Hanushek, E. A. & Rivkin, S.G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review: Papers & Proceedings* 100 (May 2010): 267–271.

Harris, D. N. (2009). Would Accountability Based on Teacher Value Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives. *Education Finance and Policy*, *4*(4), 319–350.

Ho, A. D. (2008). The problem with “proficiency”: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*. 37(6), 351-360.

Holland, P. W. (2002). Two Measures of Change in the Gaps Between the CDFs of Test-Score Distributions. *Journal of Educational and Behavioral Statistics*, Vol. 27(1), 3–17.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91-114.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER working paper*. Retrieved from http://www.nber.org/papers/w14607.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Education Finance and Policy, 6(1), 18–42.

Koedel, C., & Betts, J. (2010). Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. *Education Finance and Policy*, *5*(1), 54–81.

Koedel, C., & Betts, J. R. (2009). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Working Paper*.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4), 143-156

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, *25*(3), 287.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007) Bayesian methods for scalable value-added assessment. *Journal of Educational and Behavioral Statistics*. Vol 32(2), 125-150.

Markus, K. & Borsboom, D. (2011). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, 1, 1-11.

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

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2009). Turning student test scores into teacher compensation systems. Rand Research Report.

OECD. (2008). Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools.

http://www.oecd.org/document/54/0,3746,en_2649_39263231_41701046_1_1_1_1,00.html.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

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.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid measure? (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Spellings, M. (November 18, 2005). Secretary Spellings Announces Growth Model Pilot, Address Chief State School Officers' Annual Policy Forum in Richmond. U.S. Department of Education Press Release. Retrieved August 7, 2006 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html.

Stenner, J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are lexile text measures? *Journal of Applied Measurement*, 7(3), 307-322.

Stenner, J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: some corollaries. *Journal of Applied Measurement*, 11(3), 244-252.

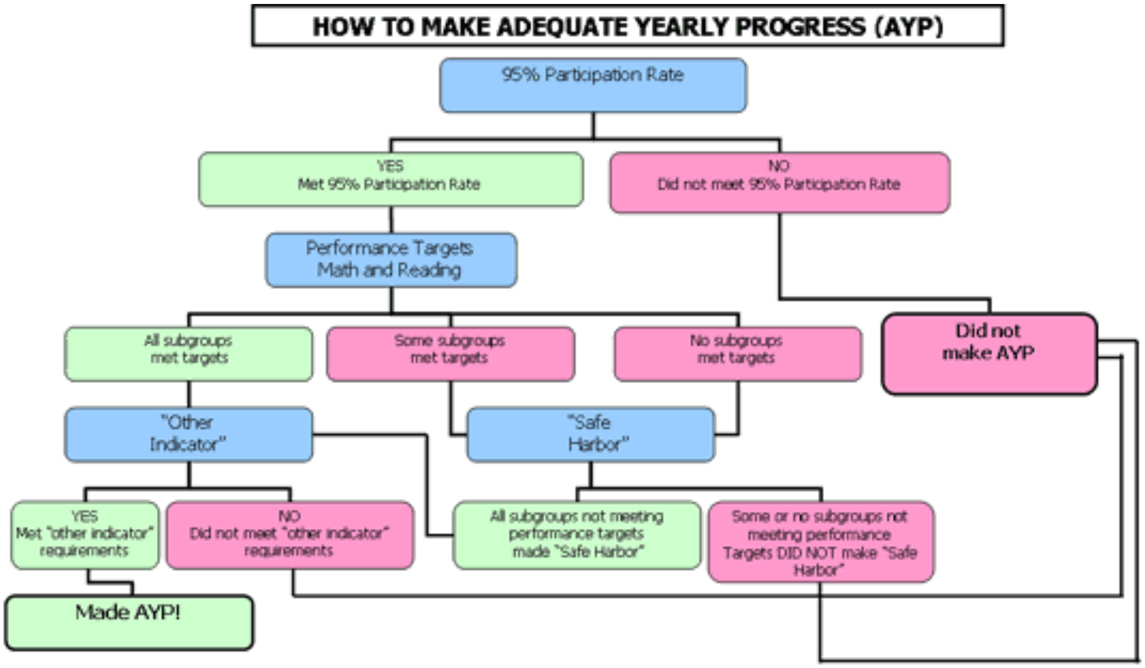
Stocking, M. L. and Lord, F. M. (1983) Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201-210.

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, 113(485), F3–F33.

Wright, P. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS White Paper. http://www.sas.com/resources/whitepaper/wp_16975.pdf

Appendix

Figure A-1. A Flow Chart for AYP Decisions in Colorado


HOW TO MAKE ADEQUATE YEARLY PROGRESS (AYP)
95% Participation Rate
YES Met 95% Participation Rate
NO Did not meet 95% Participation Rate
Performance Targets Math and Reading
All subgroups met targets
Some subgroups met targets
No subgroups met targets
Did not make AYP
"Other Indicator"
"Safe Harbor"
YES Met "other indicator" requirements
NO Did not meet "other indicator" requirements
All subgroups not meeting performance targets made "Safe Harbor"
Some or no subgroups not meeting performance Targets DID NOT make "Safe Harbor"
Made AYP!