Invitational Research Symposium on
Through-Course
Summative Assessments

SUPPORTING GROWTH INTERPRETATIONS USING THROUGH-COURSE ASSESSMENTS

Andrew Ho

Harvard Graduate School of Education

March 2011

Center for K – 12 Assessment
& Performance Management at ETS

Supporting Growth Interpretations Using Through-Course Assessments

Andrew Ho
Harvard Graduate School of Education

Executive Summary

Through-Course Assessments and Growth

One of the promises of the through-course assessment model is the support of inferences about student growth, both over the course of the academic year and toward career and college readiness in the future. The distinguishing features of through-course assessments, particularly the increased number of assessment time points and the relevance of these time points to the curriculum, are particularly well suited to support student growth inferences.

SMARTER Balanced Assessment Consortium (SBAC), Partnership for the Assessment of Readiness for College and Careers (PARCC), and Dual Goals for Growth

I describe dual goals of the SMARTER Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Careers through-course assessment proposals as they relate to growth: (a) to use multiple waves of assessment results to determine adequate student growth toward career and college readiness and (b) to use multiple waves of assessment results to support and improve teacher instruction and student learning. I refer to these as an accountability goal for growth and a learning goal for growth, respectively.

Viewing Through-Course Assessments in Terms of Current State Growth Models

Existing state growth models highlight tradeoffs that through-course-assessment-based growth models will face. Two classes of state growth models known as prediction and trajectory models have straightforward extensions to the through-course context.

Stark Contrasts Between Prediction and Trajectory Models

There are stark and unarticulated contrasts between the implications of predication and trajectory models. The prediction model is more accurate at predicting future outcomes such as career and college readiness; however, it can result in severely distorted incentives. Most distressingly, prediction models lead to *inertial inferences,* whereby initially low-scoring students can never make adequate growth and initially high-scoring students can decline freely. In contrast, trajectory models are susceptible to gaming by artificially lowering early scores, thus inflating subsequent trajectories. However, trajectory models align well with student and teacher conceptions of growth over time.

Recommendations

Recommendations for the consortia follow.

Recommendation 1

The consortia should evaluate summative growth models not only by the inferences they support but also by the incentives they create. Prediction models lead to inertial inferences and disincentivize improvement for consistently high- and consistently low-scoring students. Trajectory models may be gamed by decreasing early scores. Neither results in desirable responses on its own.

Recommendation 2

Prediction models make accurate predictions about students' future attainment of standards. However, if used, it should not be advertised as a growth model but as a prediction model. The model does not support pedagogically useful inferences about student growth.

Recommendation 3

The consortia should consider a compromise between the trajectory and prediction models in the form of a simple and transparent weighting scheme for the through-course assessments through the academic year, with relatively low but positive weights on early assessments and higher weights on later assessments. These would not support growth inferences as much as moderate incentives.

Recommendation 4

To achieve the full formative potential of growth inferences, vertical scales should be embraced. Incorporation of a common scale, even acknowledging imperfections for some subject domains and scaling designs, can reorient pedagogy toward progress over time.

Supporting Growth Interpretations Using Through-Course Assessments

Andrew Ho
Harvard Graduate School of Education

One of the promises of the through-course assessment model is the support of inferences about student growth, both over the course of the academic year and toward career and college readiness in the future. Growth inferences serve a variety of educational stakeholders. Teachers and students may plan instruction in response to past and predicted growth trajectories, and administrators and policy makers may use growth toward career and college readiness to defend the relevance and stringency of accountability models. The distinguishing features of through-course assessments, particularly the increased number of assessment time points and the relevance of these time points in the curriculum, are particularly well suited to support student growth inferences.

The policy context for this paper is the promise of widespread implementation of through-course assessment models by two consortia of states: the 31-member-state SMARTER Balanced Assessment Consortium (SBAC) and the 26-member-state Partnership for the Assessment of Readiness for College and Careers (PARCC). The SBAC proposal to the Race to the Top competition, submitted on June 23, 2010, emphasized growth in its executive summary, particularly assessments for "tracking and analyzing the progress towards college and career readiness of individual students" (Washington State & SMARTER Balanced Assessment Consortium [SBAC], 2010, p. 2). More recent details suggest that scales will be designed to support growth: "The scores for each of the through-course components . . . are on the same scale as the comprehensive end-of-year assessment for the purpose of monitoring student success as the year progresses" (N. Doorey, personal communication, December 12, 2010). The

PARCC consortium, whose emphasis on college and career readiness is evident in its very name, likewise uses its proposal summary to emphasize growth inferences such as "on track" and "meaningful progress," and there is a similar focus on benefits to educators, who may "use the results from the assessments to inform instructional decisions throughout the school year" (Partnership for Assessment of Readiness for College and Careers [PARCC], 2010, p. 3). At this time, these proposed emphases on growth are priorities identified in consortia proposals and do not represent decisions that the consortia have made. The purpose of this paper, therefore, is to frame and evaluate growth reporting and accountability options that arise from these priorities.

In this paper, I begin with an important distinction between two seemingly harmonious keystones of the SBAC and PARCC through-course assessment proposals: the accountability goal for growth, supported largely by a theme of progress toward career and college readiness, and the learning goal for growth, where growth trajectories inform instructional decisions and support student learning. My purpose is to contrast two growth models—currently in use in multiple states under the auspices of the Growth Model Pilot Program (GMPP; U.S. Department of Education, 2005)—in terms of their compatibility with these accountability and learning goals. These two growth models are already being used by states as growth models in an end-of-year assessment context, and their extensions to a through-course assessment model seem straightforward and promising. However, research is beginning to demonstrate that the contrasts between the models are not sufficiently captured by juxtaposed descriptions alone (Hoffer et al., 2011). Indeed, the trade-offs between these two accountability models are surprising and stark, extending from the incentives they set to their authenticity as models that support growth inferences.

I extend previous work by identifying these trade-offs in the context of a through-course assessment model, where I argue that the trade-offs between accountability goals and learning goals become even more vivid. One accountability model aligns well with a learning model for

growth but has unsatisfactory features for accountability purposes, and the other accountability model functions well for accountability but is poorly aligned with teacher and student conceptions of student growth. These trade-offs are not abstract. Using one accountability model will result in a quantifiable decline in the accuracy of predictions for career and college readiness. Both accountability models have features that create distorted incentives easily recognized as absurd in a classroom context. Simply put, these growth models are poorly aligned with growth goals.

I proceed by contrasting accountability and learning goals for growth. I continue on to describe the two widely used accountability models that, at face value, seem similar in their support of two of the primary goals of the PARCC and SBAC through-course proposals: to support teaching and learning and to support inferences about career and college readiness. This presentation leads naturally to the stark contrasts and trade-offs between the models. In response to these trade-offs, I conclude with a series of recommendations for the design of through-course accountability models for supporting growth interpretations.

Testing for Accountability; Testing for Learning

The contrast between accountability and learning goals will sound familiar to those with a background in the assessment literature. I will briefly describe the general contrast before describing the specific contrast in the context of growth. Assessments serve multiple purposes, and the tension between assessments for learning and assessments for accountability is well documented. Wilson (2004) described a *control-validity dimension*, where he argued that large-scale assessments have become preoccupied with standardization, replication, and control over instructionally relevant and valid tasks. Black and Wiliam (1998) reviewed the literature to show the considerable impact of formative assessments on learning gains, taking care to highlight the differences between formative and summative assessments. There is also a parallel under the heading of achievement goal motivation, particularly the contrast between mastery (formative) and performance (summative) goals and their impact on student motivation (Ames, 1992;

Covington, 2000). The threats to student learning when tying stakes to aggregate scores are documented in arguments for performance assessments (Lane & Stone, 2006) and can occasionally be quantified by inflated gains in scores on assessments that help in making high-stakes decisions (Koretz & Hamilton, 2006).

The history of accountability metrics is also rife with unintended consequences and distorted incentives. As a recent example, the proficiency metric has reached the peak of its prominence under the 2002 reauthorization of the Elementary and Secondary Education Act, commonly known as No Child Left Behind (NCLB). The proficiency metric is cut-score based and only counts the numbers of students on either side of a single cut score. As a result, the metric is insensitive to individual student progress and declines that do not cross the cut score. This insensitivity has been criticized (Ho, 2008; Linn, 2003; Rothstein, Jacobsen, & Wilder, 2006), and there is limited evidence suggesting that the resulting incentives are short-changing very low- and very high-scoring students (Booher-Jennings, 2005; Diamond & Spillane, 2004; Neal & Schanzenbach, 2007). In this paper, I argue for an analytic lens that considers not only the statistical properties of an accountability model and the policy rhetoric surrounding it, but the actual incentives the models set for teachers, administrators, and policy makers.

The SBAC and PARCC proposals embraced the contrast between assessment for learning and assessment for accountability, with the explicit intention of achieving both goals. The SBAC proposal followed the principle of *responsible flexibility* and proposed the use of a range of item types in a computerized adaptive testing system to report student scores on a standardized scale. This would ostensibly allow for tests with more flexibility, including teacher scoring and development of performance tasks (Washington State & SBAC, 2010). The PARCC (2010) proposal similarly emphasized sophisticated items and performance tasks that "will help model effective classroom instruction." The PARCC proposal hewed closer to a unidirectional measurement-informed instruction model than the more bidirectional SBAC proposal, which emphasized teacher-designed measurements as well as measurement-informed teaching.

However, both proposals suggested that through-course assessments can serve both learning goals and accountability goals with innovative assessments, short turnaround times, and more frequent administrations.

The companion papers in this symposium address some of the challenges and possibilities of the through-course assessment model in achieving accountability and learning goals. The specific purpose of this paper is to identify intrinsic trade-offs between two accountability models as they function to support growth inferences for accountability and learning. Unlike the general contrast, which crosses many disciplines within education, I define the growth contrast more narrowly. An accountability model for growth uses multiple waves of assessment results to determine adequate growth for a student. In contrast, a learning model for growth uses multiple waves of assessment results to improve teacher instruction and student learning by, for example, improved targeting of instructional resources or enhancement of students' metacognitive processes. By extending the inference from one test administration to many, and from snapshot status to progress over time, I intentionally restrict the contrast between models to those where growth inferences are the target. In this light, the trade-offs between models for growth become particularly well defined.

Accountability Models for Growth

The SBAC and PARCC proposals can be seen as attempts at correcting the missteps of accountability models of the past. In particular, NCLB mandated individual student growth only inasmuch as students are required to be *proficient,* that is, scoring above a judgmentally determined cut score. This accountability model is wholly insensitive to individual student growth unless a student changes status from nonproficient to proficient or the reverse. Furthermore, there was little attention to the vertical articulation of the proficient standard across grades. As the definition of *proficiency* varied across grades, a student's so-called progress from nonproficient in one grade to proficient in the next grade had little connection to growth. Ho, Lewis, and Farris (2009) identified scenarios in which students can decline in terms

of relative standing and nonetheless progress from nonproficient to proficient. This poor alignment between school function (growth) and school accountability (in this case, status) threatened undesired responses, including the sanctioning of low-status, high-growth schools and inattention to students with low probabilities of changing status.

A recognizable federal response to this misalignment began in 2005 with the introduction of the GMPP (U.S. Department of Education, 2005). The GMPP allowed approved states to incorporate individual student growth into their NCLB accountability calculations. Fifteen states were authorized under the program, and many more have proceeded with similar models in accordance with Race to the Top criteria. A so-called bright-line principle of the GMPP was that students had to be on track to proficiency by 2014; that is, the central theme and goal of NCLB had to be preserved. The SBAC and PARCC proposals have embraced a very similar on-track principle, following the Race to the Top request for proposals: to ensure that students are on track to career and college readiness. However, there is no single, straightforward approach to using through-course assessment data to determine progress toward career and college readiness. In this section, I overview two models that are intuitive, compelling, and effective, and I demonstrate stark differences in their classification approaches that will have an impact on the validity of growth inferences and uses. These models can incorporate multiple scores from a single academic year, as in the through-course model, or multiple end-of-year scores, as under current state growth models, or a combination of both. In addition, these models are flexible to definition of the ultimate target, whether it is some definition of career and college readiness or a general notion of "grade-level proficiency" as defined by future proficiency cut scores.

As a point of contrast, it is worth mentioning that I do not consider these growth models in the traditional statistical or psychometric sense. These models create complex incentive structures and are primarily concerned with improving student and school outcomes, not with understanding the functional form of growth trajectories (e.g., Rogosa & Willett, 1985) or even,

for this application, with trying to ascribe growth to various value-added causal effects (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). Instead, these are better described as *accountability models for student growth*: They may use statistical methods, but they are overlaid with numerous policy decisions and are, more essentially, tools of policy, designed to classify students and schools and thereby change behavior and improve educational outcomes. The desired contrasts are not of a statistical nature, where bias, standard errors, and cross-validated recovery of parameters might be criteria of interest; instead, the desired contrasts are of a practical nature: Who is classified as what by which model? What are the hidden dependencies that might diminish transparency and alignment? Although I may use the terms *accountability model* and *growth model* interchangeably, I am referring to the accountability functions of the latter and not simply their statistical features.

Two Contrasting Accountability Models for Growth

As a hook, let us consider two student score trajectories over four through-course assessments in an academic year. The assumed context is an SBAC approach whereby through-course assessment scores can be located on a common score scale, for example, the end-of-year test score scale. Figure 1 shows two student trajectories over the academic year on a hypothetical 100-point scale. Student A starts with low score of 30 and gains 10 points over each assessment period. Student B starts with a high score of 80 and declines 10 points over each assessment period. An arbitrary threshold mimicking proficiency in an NCLB context is noted for interpretation, although it is not essential to the example at this point.

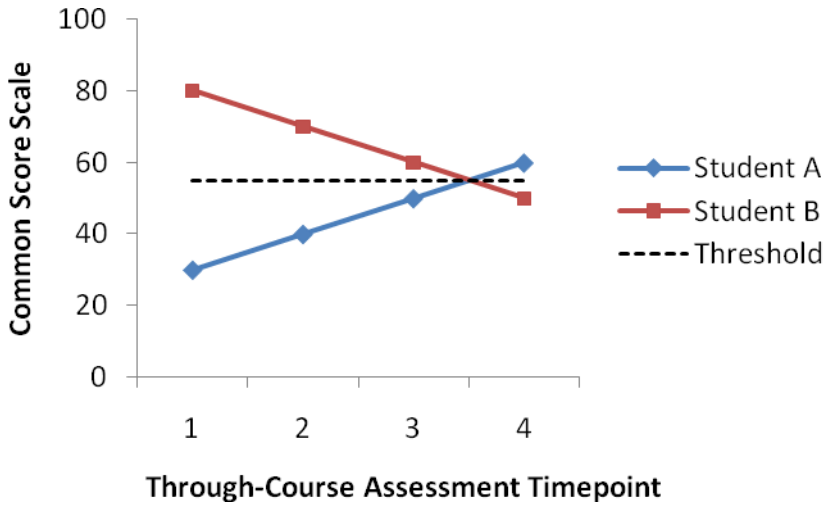

Figure 1. Two student trajectories in a through-course assessment model with a common score scale.

In a traditional end-of-year, status-based accountability model like NCLB, Student A would be considered proficient and Student B would be considered not proficient, based solely on the score on the end-of-year assessment at Time 4. The addition of the through-course assessments at Times 1, 2, and 3 provides us with the promise of enhanced inferences about student growth, both to support our accountability model, as we argue for the meaningfulness of the student's trajectory in terms of career and college readiness, and to support our learning model, as the student growth trajectories give a clear indication to teachers and the respective students that Student B is declining in achievement and Student A is improving. At a glance, we would expect an accountability model for growth to be consistent with the end-of-year inference: Student B is clearly not on track for career and college readiness, and Student A is making adequate growth.

The punch line is that one very popular accountability model known as a *projection model,* or more appropriately, a *prediction model* or *regression model,* will disagree with the seemingly obvious observation that Student A is the one making adequate growth. This model

will instead suggest that Student B has a greater likelihood of being on track to career and college readiness. This would seem to be absurd, and one might think that any model that made such a claim should be dismissed from consideration. However, among the reasons that this model has been widely adopted is this: Student B is, in fact, more likely to be on track to career and college readiness than Student A.

To understand how such a situation arises, let us formalize the contrast between the two models of interest. The first is a *trajectory model,* one that is visualized in Figure 1 as a plot of student scores on a common score scale over time. In the case of Students A and B, the trajectories are perfectly linear, but a general trajectory can be defined for each student by the best-fit regression line for each student's scores regressed on time. This linear trajectory can be extended into the future, and determinations about whether students are on track to reaching a particular threshold in the future, be it proficiency accorded by a particular grade or a college–career readiness benchmark on the score scale, follow in a straightforward fashion. Applying a trajectory model to the data in Figure 1 leads to a conclusion that Student A is making progress, that Student B is declining, and that if anyone were on track to some high standard in the future, it would be Student A.

The second model is a *prediction model.* It is also regression based, but the target of inference is not an individual student's linear trajectory, as in the trajectory model, but a general prediction equation, applicable to all students, that takes each student's four test scores and uses them to predict some outcome such as career and college readiness in the future. The model prediction can be expressed as $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4}$. Here $\hat{Y}_i$ is some predicted outcome in the future for student i and is obtained simply by plugging in that student's four test scores, $X_{i1}, \ldots, X_{i4}$. The regression coefficients (βs) are estimated by an ordinary least squares regression using a previous cohort of data where the future outcome *Y* is observed along with through-course assessment scores. Prediction models and trajectory

models are both accountability models for growth that are widely used for state accountability purposes (Hoffer et al., 2011).

The disagreement about whether Student A or Student B is making adequate growth can be seen in the functional form of the prediction model, which does not reference the order in which the test scores occur. From this vantage, we can see that Student B has, on average, higher scores than Student A, but we are blind to the order in which the scores occurred. In such a situation, the logic of a prediction model, where higher scores predict higher future outcomes, seems more compelling. Furthermore, owing to the well-known usefulness of regression models for the purpose of prediction, the appropriately named prediction model does just that, and its accuracy of predicting future status, such as career and college readiness, is superior to that of trajectory models. The crux of the contrast lies in the trade-off between accuracy in future prediction, a highly desirable feature of an accountability model that rests on an argument for career and college readiness, and instructional relevance, where Figure 1 would seem to any educator and student to clearly indicate the growth of Student A over Student B.

A Framework for Understanding Growth Contrasts

Let us imagine a simple through-course assessment scenario that might arise in a SBAC or PARCC classroom. A student is midway through her academic year and has been exposed to two through-course or interim assessments. For simplicity, let us assume that the student has two scores: a score at time $g-1$ and a score at time *g*. The desired inference for accountability is whether this student is making adequate growth. The PARCC proposal and, to a lesser degree, the SBAC proposal have both emphasized benchmarking by career and college readiness standards. Let us imagine that this affords us a series of cut scores above which a student is ready and below which a student is not ready. As the student progresses through a series of through-course assessments, each of these scores—let us call them X_{ig} for student *i*

at time *g*—can be compared with these cut scores, c_{pg}. If the student scores above the cut, she has met the readiness standard.

Just like the proficiency standard under NCLB, this readiness standard does not incorporate growth information. For growth inferences, an additional classification becomes relevant: whether a student is on track to career and college readiness. A student may be below the standard but nonetheless on track to reach the standard in the near future. Likewise, a student may be above the standard but on track to fall below the standard in the near future. In a 2009 paper, Ho and colleagues introduced a graphical framework for understanding the effects of cut-score selection on growth models. The framework was adapted and extended as part of a U.S. Department of Education evaluation of the GMPP (Hoffer et al., 2011). I extend the framework further here, incorporating the context of through-course assessment and paying particular attention to a single contrast of interest.

The graphical framework is introduced in Figure 2. The horizontal and vertical axes of Figure 2 represent the initial time $g-1$ and current time *g* student scores, respectively. The scales are standardized to the *z* scale for interpretability: A 0 represents an average score, and a –1 represents a score 1 standard deviation unit below the mean. Students with increasing standardized scores lie above the diagonal, and students with declining standardized scores lie below the diagonal.

The scatter plot is purposefully centered in the low-scoring region of the distribution, where most state proficiency cut scores are set. The cut score is assumed to be at the same relative level across time $g-1$ and *g*, at the score of –0.5 for each year. This results in around 70% of students meeting the standard at each administration, a typical proficiency rate found in practice (“State of the States,” 2010). If student scores were distributed normally with a moderately high correlation, as expected, the scores would cluster loosely around the diagonal drawn on the figure and be most dense toward the upper right corner of the graph, near the center of the bivariate distribution, (0,0).

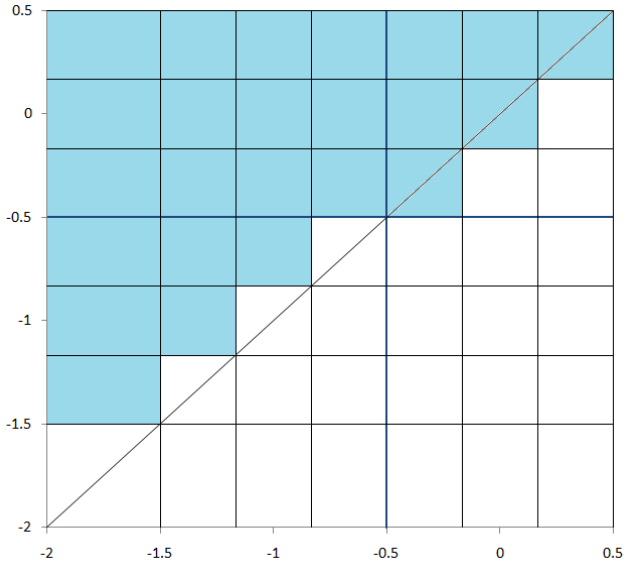

Figure 2. A categorical growth model identifies students who have gained or maintained a category from time g – 1 to time g. Shaded regions indicate bivariate score regions where students will be classified as having made adequate growth on track to a standard. The cut scores are bolded, and the diagonal representing unchanging relative scores is displayed for reference.

The usefulness of Figure 2 is that it allows visualization of the areas where students are classified as having made adequate growth. One such model is highlighted in blue in Figure 2. Categories along the diagonal but below the cut score are not highlighted, but categories on the diagonal above the cut score are highlighted. This corresponds to a model that specifies that below-standard students must gain a category but above-standard students need only maintain their category. The categorical nature of the model is clear in the gridlike nature of the boundaries. In the next sections, I contrast the shaded patterns of the target models of interest.

The Trajectory Model

A common model in the GMPP program, and a model quite similar to that implied by the SBAC proposal, is the trajectory model, a *momentum model* that gives credit to below-

standard students who have nonetheless made sufficient recent gains that they identify a trajectory to proficiency by a certain time horizon *h*, where *h* is some set number of time units in the future. For a student's current and past assessment scores, X_{ig} and $X_{i,g-1}$, this might be expressed as follows: If $3*(X_{ig} - X_{i,g-1}) + X_{ig} \geq c_{p,g+3}$, then student *i* is on track. In other words, if the gain over the past test, $X_{ig} - X_{i,g-1}$ $X_{ig} - X_{ig-1}$, is extended for three time units, and the student has met the standard at that point, then the student is on track. Alternatively, this could be expressed: If $X_{ig} - X_{i,g-1} \geq (c_{p,g+3} - X_{i,g-1})/4$, then student i is on track. In other words, if the distance between the initial score and the target in the future is divided by 4, and the gain over the past year exceeds it, then the student is on track to proficiency. Of course, both expressions are algebraically equivalent.

Figure 3 begins to reveal the full potential of the graphical framework, as it allows the identification of the classification approach of the trajectory model in contrast with that of the previous illustrative categorical model. For this version of the trajectory model, the equation can be simplified to slope-intercept form, $X_{ig} \geq (3/4)X_{i,g-1} + (c_{p,g+3}/4)$. By further assuming that the model applies to both below-standard students (on track in 3 years) and above-standard students (preventing declines below standard), the trajectory model extends past the cut score to the upper right of the scatter plot. If we assume for simplicity that the cut score $c_{p,g+3}$ exists at the same relative level in future time, the inequality can be plotted in a straightforward fashion. The highlighted region is similar to Figure 2, but smoother. This contrast allows visualization of the previous categorical model as a discrete approximation of the trajectory model.

Figure 3 also allows the visualization of an interesting feature of trajectory models (and their categorical counterparts). For any given Time 2 score, X_{ig}, a student would be better off with a lower Time 1 score, $X_{i,g-1}$. In Figure 3, this can be seen by picking any horizontal line

representing a common Time 2 score and observing that shifting left, or decreasing the Time 1 score, allows the student to approach the highlighted region. This seems jarring from an accountability standpoint but is a logical extension of a trajectory-based model. For a particular current-grade score, the lower the previous grade scores are, the greater the estimated trajectory will be.

In practice, the distorted incentives arising from this model are diminished somewhat by the two additional layered provisions. First, there is an implementation of a straight proficiency cut score at Time 1, represented by the vertical line at $c_{p,g-1}$. Gaming a trajectory model by deflating initial scores would require not only understanding of trajectory model features and healthy cynicism but also moderate assurance that the student would not otherwise be proficient. Second, there is typically a caveat that a student must actually reach a target in a specified amount of time. Deflating initial scores by actively withholding instructional resources risks the eventual future attainment of standards within the required time frame.

The Prediction (Projection) Model

The prediction model aligns closely to the language of the PARCC proposal, which focuses more than the SBAC model on the prediction of career and college readiness in the future. The model uses scores from a previous cohort of students to estimate the parameters of a prediction equation. This prediction equation is then used to predict the future scores of current students. Given scores *Y* from a previous cohort, a regression equation of the form $\hat{Y}_{i,g+h} = b_0 + b_1 Y_{i,g-1} + b_2 Y_{ig}$ can be estimated. Substituting current scores for the previous cohort, a future score, $\hat{X}_{i,g+3} = b_0 + b_1 X_{i,g-1} + b_2 X_{ig}$, can be predicted. The decision rule is then, simply, that if $\hat{X}_{i,g+h} \geq c_{p,g+h}$, then the student is on track. The coefficients b_0, b_1, and b_2 cannot be estimated from current data because the variable $X_{i,g+3}$, *h* time units in the future, is unavailable.

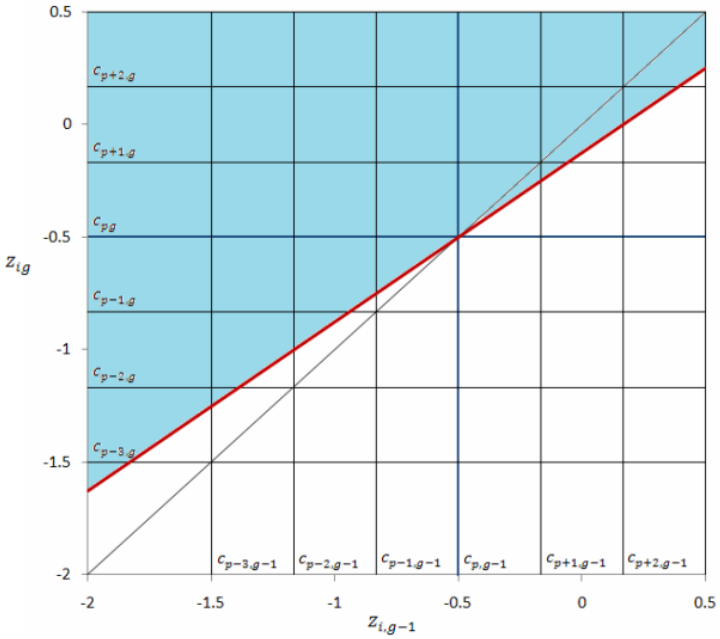

Figure 3. A trajectory growth model identifies students whose gains from time g – 1 to time g set them on a trajectory to or above the standard by some future time horizon. The shaded region indicates bivariate scores whereby students will be classified as having made adequate growth on track to proficiency. The cut scores are bolded, and the diagonal representing unchanging relative scores is displayed for reference.

To increase the likelihood of accurate predictions, states with GMPP-type prediction models use many more grades and also multiple subjects to estimate prediction equations, and school-level centering is also employed. Extensions of this model to include teacher-level variables have been characterized as "value-added models" because of the desired inferences for teacher-level effects (McCaffrey et al., 2004). However, the essential features of the prediction model are captured well by the simple, two-predictor equation.

When there are only two predictors and variables are standardized, the two statistics b_1 and b_2 can be estimated solely from the intercorrelations between times $g-1$, g, and

$g+h$ in the previous cohort. Following the usual matrix derivations of ordinary least squares regression coefficients, one can obtain the following:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & r_{g-1,g} \\ r_{g-1,g} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{g-1,g+h} \\ r_{g,g+h} \end{bmatrix}.$$

Plugging these estimates back into the prediction equation gives the desired inequality that identifies students as on track to career and college readiness, $b_1 X_{i,g-1} + b_2 X_{ig} \geq c_{p,g+h}$. By again assuming that the cut score $c_{p,g+h}$ exists at the same relative level, this inequality can be plotted as an area on the bivariate scatter plot. Perhaps surprisingly, for most realistic cross-test correlations in reference cohorts, the prediction model's target region looks similar to that shown in Figure 4: a marked contrast to the trajectory model. The coefficients that define the line in Figure 4 were plotted from real-world, cross-grade correlations from a mid-sized state where $h=3$. Cross-grade correlations from alternative grades and subjects—and even cross-grade correlations gathered from technical reports of other state tests—do not change the most salient feature of the line: its negative slope. Figure 4 retains the trajectory model line for contrast but removes the categorical model lines for clarity.

Figure 4 reveals stark contrasts in the classification approaches of widely used growth models. Prediction models, easily the most statistical of the three presented models in their incorporation of covariances and multiple predictors, do not explicitly model growth over time. The salience of the current time over the previous time is only represented inasmuch as adjacent-test correlations may be greater than distal-test correlations. In contrast, trajectory models explicitly require growth over time or, for above-standard students, discourage significant declines. The inattention of the prediction model to the ordering of scores may seem jarring to policy makers and educators, even as it alludes to the nature of statistical prediction: High scores predict high scores, regardless of their order.

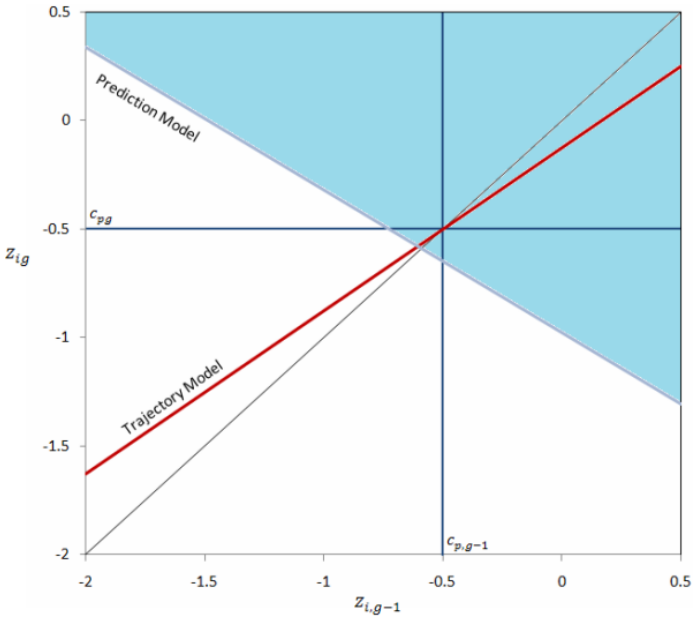

Figure 4. A prediction growth model identifies students whose current and past scores predict a future standard according to a prediction equation. The shaded region indicates bivariate scores where students will be classified as having made adequate growth on track to the standard. The cut scores are labeled, and the diagonal representing unchanging relative scores is displayed for reference.

The distinction between prediction models and trajectory models is nonetheless confounded in the literature, where prediction models are often described as "projection models" (e.g., Dunn & Allen, 2009), making them sound similar in metaphor to trajectory models. Figure 4 demonstrates that the contrasts extend from the model classification approach to the function of the models as policy tools. Owing to its statistical approach, the projection model can be expected to outperform the other models in terms predicting whether students have actually met the standard by the eventual time horizon. However, the educational desirability of crediting previously high-scoring decliners may be less than recognizing previously low-scoring gainers.

From the perspective of incentives, prediction models also contrast with trajectory models by demanding high scores at both time points. That is, for a given Time 2 score (across a horizontal line), it is better to have a higher Time 1 score, and for a given Time 1 score (along a vertical line), it is also better to have a higher Time 2 score. This seems more natural than the incentives set up by the trajectory model even as the instructional effect is counterintuitive. Additional provisions, such as disallowing previously proficient decliners, reduce this model's impact to a small central triangle defined by the two proficiency cut scores at Times 1 and 2 and the prediction model cutoff. This leads to the unsurprising finding that prediction models ultimately have little impact on school classification decisions in a proficiency-centered framework (Hoffer et al., 2011). Without a shift outside the proficiency-centered framework, there is very little of a growth model acting at all.

Framework Summary

The bivariate framework is useful in deriving the link between models and their classification approaches, but it is somewhat complex and not immediately interpretable without moderate exposition. A more accessible representation contrasts different models' approaches to the minimum adequate growth, conditional on initial status. This required growth representation is readily interpretable as the minimum growth expectations for low-, mid-, and high-scoring students. It may be derived directly from the bivariate scatter plots by subtracting out the $z_{ig} = z_{i,g-1}$ main diagonal.

Figure 5 shows the resulting representation for the models presented in the previous sections. The vertical axis represents the gains required by each model. The horizontal axis represents the initial status of the students, with low-scoring students on the left and high-scoring students on the right. The cut score is again highlighted for reference. The trajectory model requires less than half a standard deviation unit of gains for low-scoring students with *z* scores less than 2. Students scoring at the cut score require no gains, and small declines are

allowed for higher scoring students: up to half a standard deviation of decline for students with *z* scores greater than 1.

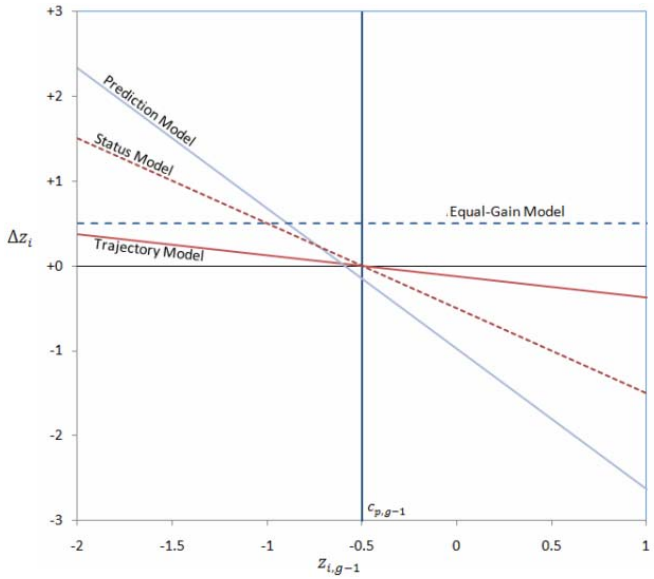

Figure 5. Growth-to-standard models contrasted with a status model and an equal-gain model in terms of required growth by initial status.

The most dramatic feature of Figure 5 was forecasted by Figure 4: the substantial difference between the prediction model and the others. Although the precise parameters depend on the cross-test correlations, Figure 5 shows that prediction models are much more demanding of low-scoring students and much more lenient for high-scoring students. The slope of the prediction model line is, in fact, almost 7 times steeper than that of the trajectory model line, meaning that lower scoring students require around 7 times the gains to meet the standard under the prediction model and higher scoring students are allowed around 7 times the decline. This contrast is, again, a natural extension of regression models with similarly and highly correlated predictors.

This representation also reveals the possibility of alternative and contrasting models. As previously noted, the trajectory and prediction models were designed to be growth-to-standard models to comply with Race to the Top implications. They thus require cut scores and the

support of an inference about a student being on track. In contrast, Figure 5 demonstrates that a growth model need not be referenced to a cut score. The simplest such growth model, here described as an *equal-gain model,* requires equal growth in scale score points for all students, regardless of their location in the distribution. The equal-gain model is deeply scale dependent and would have unfavorable properties if scores were standardized as in this illustration, but the contrast is illustrative.

Figure 5 also allows visualization of a model that is not generally thought of as a growth model: an NCLB-type status model. If cut scores over time can be located on the same vertical scale, the status model can be easily represented as a growth model: It is, in fact, a trajectory model with a time horizon of zero. Students scoring below the cut score require growth to the cut score by the next time point, and students scoring above the cut score must similarly avoid declining beneath that cut score at the next through-course assessment. The juxtaposition of the status model and the prediction model emphasizes the uniqueness of the prediction model as a growth model. It is, in some senses, more stringent for low-scoring students than the NCLB status model it replaces.

Implications for Through-Course Assessment Models

The previous sections have demonstrated a stark and consequential contrast between two attractive models for growth interpretations arising from through-course assessments. The prediction model may be best for identifying students on track to targets, whether career and college readiness or "grade-level proficiency." However, low- and high-scoring students will appear to have inertial tendencies that allow for impossibly high or absurdly low expectations. The trajectory model fits better with notions of progress but sacrifices accuracy of future predictions and allows the gaming by deflation of initial scores to inflate future trajectories. As an illustration, we might consider two entirely plausible scenarios. By midyear, Student A has not performed well on the two interim assessments administered so far. By the prediction model, Student A can do nothing to be on track by year's end, not even scoring perfect scores

on the remaining assessments. Student B, meanwhile, has performed very well—so well, in fact, that she may score absolute zeros on the subsequent assessments and still be considered on track by the prediction model. Nonetheless, this prediction model is very accurate, and the likelihood that Student A and Student B are correctly classified is very high.

Interestingly, the SBAC and PARCC proposals seem to point each consortium toward a different model, if only tentatively. The SBAC through-course approach seems to desire the location of interim assessment scores on a common scale, a procedure that will afford the kind of vertical scale that a trajectory model requires. The PARCC approach, as suggested by its name, is committed to the prediction of readiness for college and careers and may thus be more inclined to select a model that maximizes classification accuracy. In practice, a concession toward the distorted incentives hidden in both models may be the best compromise. Looking at Figures 4 and 5, the trajectory model's negative weight on initial scores may lead to initial score deflation, and the prediction model's positive weights lead to inertial tendencies. The compromise, a status model, as seen in Figure 5, has a zero weight on initial scores and is not a growth model at all. A compromise toward prediction and positive weights for all scores leads to a steeper negative slope than the status model in Figure 5. A compromise towards intuition and instructional relevance leads to a less shallow negative slope than the status model in Figure 5. The greater the summative stakes on the model, the more the steeper slope might be preferred, although not to the extent that inertial tendencies seem impossible to overcome. The lower the summative stakes on the model, the more the trajectory model's approach will be preferred.

My recommendations for the consortia are fourfold. First, the consortia should evaluate growth models not only by the inferences they support but also by the incentives they create. The long history of accountability metrics reveals an alarming tendency to sacrifice the actual incentivizing function of an accountability model in order to frame it with desirable rhetoric. There needs to be adequate transparency in these models to reveal the incentive structures

and ensure that accountability models for growth actually incentivize growth for the target population.

Second, the prediction model should be advertised as what it is: a model for prediction that only responds to the predictive utility of its inputs. As a point of fact, any variable may be used as a predictor, regardless of whether it is a test-score variable or has any substantive relevance to the prediction of future scores at all. As a consequence, prediction models are much less growth models, as their functioning in the framework presented here makes transparent. As such, the prediction and classifications from the model should only be used cautiously if at all to guide instruction. Such a system, whereby some students would be predetermined to meeting or failing the standard halfway through the year, would be ripe for gaming and misuse, and it is clear that these incentives runs counter to the pedagogical goals of any classroom or accountability model.

Third, a compromise model or dual models functioning in parallel may be an improved approach to supporting growth inferences while making accurate predictions. The higher the stakes, the more moderated these models should be, as might be visualized in Figure 5 with a line between the status model and the prediction model. The trajectory model, as a legitimate model for growth over time, may be the model used to guide instruction throughout the year. At year's end, a prediction model may begin to be incorporated. A *straight prediction mode*l would be undesirable as it would essentially override the incentives created by the trajectory model over the year. A *masked prediction model,* in which the prediction equation is hidden, might lessen gaming but would run afoul of transparency principles and would nonetheless be fairly predictable and distort classroom incentives. A *hybrid prediction model* may therefore be a compromise. The small sacrifice to classification accuracy would be offset by recentering incentivizes towards gains over time.

Fourth, if growth interpretations in a curriculum are truly a target inference, a relatively inexpensive vertical scale with imperfect properties may be preferable to a model that does not

use any vertical scale at all. The prediction model not only distorts incentives but also functions as a black box—a counterintuitive and abstract prediction equation that, while accurate, holds little relationship to student growth over time as it is commonly understood. Incorporation of a common scale, even acknowledging limited flaws, may reorient pedagogy toward progress over time instead of neutral and often inertial prediction models. As these proposals reflect good-faith efforts toward using growth to support both accountability and learning goals, the evidence presented in this paper supports an argument for vertical scales underlying through-course assessments, without which instructionally relevant inferences about student progress over time become challenging, if not impossible.

References

Ames, C. (1992). Classrooms: Goals, structures and student motivation. *Journal of Educational Psychology, 84*, 261-271.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139-148.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42,* 231–268.

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology, 51,* 171-200.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106*, 1140-1171.

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

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

Ho, A. D., Lewis, D. M., & Farris, J. L. M. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15–26.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., et al. (2011). *Final report on the evaluation of the Growth Model Pilot Project*. Washington, DC: U.S. Department of Education.

Koretz D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education/Praeger.

Lane, S., & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 387-431). Westport, CT: American Council on Education/Praeger.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. Education Policy Analysis Archives, 11(31). Retrieved March 15, 2007, from http://epaa.asu.edu/epaa/v11n31/

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

Neal, D., & Schanzenbach, D. W. (2007). *Left behind by design: Proficiency counts and test-based accountability*. University of Chicago. Retrieved July 30, 2007, from http://www.aei.org/docLib/20070716_NealSchanzenbachPaper.pdf

Partnership for Assessment of Readiness for College and Careers. (2010, June 23). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50*, 203–228.

Rothstein, R., Jacobsen, R., & Wilder, T. (2006). 'Proficiency for all'–An oxymoron. In *Examining America's commitment to closing achievement gaps: NCLB and its alternatives*. Symposium conducted at the meeting of the Campaign for Educational Equity, New York, NY. Retrieved from http://www.epinet.org/webfeatures/viewpoints/rothstein_20061114.pdf.

State of the states: Sources and notes. (2010). *Education Week, 29*(17), 49–50.

U.S. Department of Education. (2005, November 18). *Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond* [Press

release]. Retrieved February 12, 2007, from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

Washington State, & SMARTER Balanced Assessment Consortium. (2010, June 23). *Race to the Top Assessment Program application for new grants*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf

Wilson, D. (2004). Assessment, accountability and the classroom: A community of judgment. In D. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd yearbook of the National Society for the Study of Education* (pp. 1-19). Chicago, IL: University of Chicago Press.