DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Accuracy, Transparency, and Incentives: Contrasting Criteria for Evaluating Growth Models.

Citation

Ho, Andrew D. 2012. Accuracy, Transparency, and Incentives: Contrasting Criteria for Evaluating Growth Models. Paper presented at the conference on Value Added Modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness. College Park, MD: Maryland Assessment Research Center for Education Success.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:27471536

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Accuracy, Transparency, and Incentives:
Contrasting Criteria for Evaluating Growth Models

Andrew Ho
Harvard Graduate School of Education

Ho, A. D. (2012). *Accuracy, transparency, and incentives: Contrasting criteria for evaluating growth models.* Paper presented at the conference on Value Added Modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness. College Park, MD: Maryland Assessment Research Center for Education Success

The scope and complexity of modern educational accountability models make validation efforts extremely difficult. Interpretations and uses of model results are often multitudinous, underspecified, and subject to change over time. Targets of interpretation are less individual scores than aggregations and adjustments of these scores, processed through a series of often ad hoc compositing procedures and policy judgments. The models that determine school "Adequate Yearly Progress" or teacher ratings may use statistical models but ultimately depend on a series of other decisions and support a range of uses and interpretations. They are better described as accountability models or policy models. Evaluating these models requires criteria beyond statistical bias and precision and should begin by clarifying model function and purpose.

The current state of the art of validation theory is well described by Kane's 2006 and 2013 expositions. However, there remains a practical disconnect between validity theory, which benefits from well-defined scores and clear statements about score use and interpretation, and accountability models in education policy. In policy, there is arguably a benefit to poorly defined terms like "proficiency" and "college readiness." In this landscape, clear definitions and well specified theories of action may work against consensus by providing footholds for disagreement. Ambiguity is particularly useful in U.S. federal educational policy, to allow states and local districts the flexibility to which they are historically accustomed under the 10th amendment of the constitution. Although regulatory bodies can enforce specification of theories of action through, for example, guidelines and requests for proposals, the incentives behind policy formation do not naturally result in the raw materials necessary for Kane's "Interpretation/Use Arguments" (2013, p. 14). Much of the ambiguity arises from poor specification of what "scores" in accountability models ultimately are, let alone the interpretations and uses that they support.

This is no less true when it comes to "growth." At the announcement of the Growth Model Pilot Program (GMPP) in 2005, Secretary Margaret Spellings never defined what "growth" meant (U.S. Department of Education, 2005). The announcement insisted only that models adhere to seven "bright line principles," such as "ensure that all students are proficient by 2014," and that the model "must track student progress." Subsequent guidelines from the peer review panel similarly left latitude for growth definition and model specification (U.S. Department of Education, 2006). The guidelines at the announcement of the Race to the Top competition seemed to be more explicit, and defined growth as “the change in achievement data for an individual student between two or more points in time” (U.S. Department of Education, 2009, p. 59742). However, it then continued, “a State may also include other measures that are rigorous and comparable across classrooms,” along with its motivations, “to allow States the flexibility to develop data and assessment systems” (p. 59742). This left space for GMPP models to continue and effectively took no position on the definition of “growth.”

As the final report of the GMPP described (Hoffer, et al., 2011), states took a variety of approaches to operationalizing growth. Models have continued to proliferate in the Race to the Top era. In the first part of this chapter, I review four prototypical models and demonstrate that these models operationalize growth using related but fundamentally distinguishable approaches. In the second part of this chapter, I articulate three contrasting criteria: predictive accuracy, transparency, and incentives, and I contrast the models on these dimensions. The key observation that this chapter supports is that models that excel at certain criteria are substandard at others. This reinforces the need to explicate desired theories of action and criteria early, and select models, metrics, and reporting principles that align with these. A secondary observation is that the functional unit of analysis is less a model than a metric, where an important factor is the level of aggregation at which it functions. A feature of a metric at one level of aggregation may be a flaw at another level of aggregation. Although validation of growth metrics is not the purpose of this chapter, I hope to provide the raw materials for validation by clearly articulating models and metrics and identifying criteria along which they contrast.

Contrasting Foundations Underlying Growth Models

The proliferation of growth models in a policy space constructed deliberately to allow for flexibility has led to confusion among terms and definitions. Here, I follow the general framework and nomenclature provided by Castellano and Ho’s *A Practitioner’s Guide to Growth Models* (2013a), with some liberties taken for simplicity of exposition. In their guide, Castellano and Ho attempt to be explicit about each growth model and include its aliases and statistical foundations. In addition, they articulate the primary interpretations that the model supports and the levels of aggregation at which they are supported.

Table 1, below, shows a simplified version of their framework adapted to suit this exposition. I will describe column headings and then discuss each growth model in turn.

Table 1. Growth models, their statistical foundations, and the growth metrics they support.

Model		Statistical Foundation	Growth Description		Growth Prediction	
			Student Level	School Level	Student Level	School Level
Gain-Based	Trajectory, Slope, Difference, Gain	Difference: Current Score Minus Past Score	Gain Score	Percentage of Acceptable Gains	Trajectory Model	Percentage On Track
Categorical	Value Table, Transition Matrix	Categorical. Changes in Categories.	Value of Category Change	Averaged Values of Category Changes	(Projected Category)	(Average Projected Category)
Student Growth Percentile	Colorado Model, Betebenner Model, Percentile Performance Index	Quantile Regression. Conditional Status.	Student Growth Percentile	Median Student Growth Percentile	Student Growth Projection	Percentage On Track
Projection	Regression Model, Residual Model, Multilevel Model, Hierarchical Model	Regression. Conditional Status.	Residual, "Residual Gain"	Average Residual	Projection Model	Percentage On Track

The first column of Table 1 highlights the model names as I will refer to them in this chapter. The gain-based model is an intuitive and largely straightforward model that requires a vertical scale. The categorical model is a flexible framework that considers student status in a small number (usually 4 to 9) of categories and operationalizes growth in terms of transitions between categories (Hill, et al., 2006). The Student Growth Percentile (SGP; Betebenner, 2009) model expresses growth in terms of percentile ranks using quantile regression. Finally, the projection model uses regression-based methods to define growth and make predictions.

The second and third columns list aliases and statistical foundations of the models, respectively. Aliases are essential given the rapid proliferation and casual use of terms relating to growth. The categorical model is also known as a value table or a transition matrix model. The projection model is often confused with the trajectory model, as the metaphors of projection (like a movie projector) and trajectories (like a ball through the air) are similar. However, the two are fundamentally different, practically and statistically. The statistical foundations of the gain-based model are intuitive change scores. The categorical model uses changes in categories, and SGPs and projection models use conditional status: the observed status of students compared to their expected scores conditional upon past scores.

The final four columns list the interpretations that the models support and the level of aggregation at which the interpretations are supported. Table 1 shows that a given model can support two different primary interpretations, growth description and growth prediction. This contrasts with Castellano and Ho (2013a), who describe the gain-score and trajectory metrics, for example, as separate models entirely. Growth description refers to inferences about growth up to and including the most recent data. Growth prediction refers to inferences that rely on prediction, using past growth data, about some future time point. Table 1 also shows that the metrics support interpretations at different levels of aggregation. Here, we focus only on the student level and the school level as illustrations. As I note in the introduction to this chapter, the rules of aggregation must be clearly specified before beginning a validation effort. Simply distinguishing between two levels of aggregation is hardly sufficient to represent the complex decisions that go into, for example, Adequate Yearly Progress calculations, but it does begin to capture the reality that different "scores" function at different levels and have different implications.

The metrics listed are examples and are not meant to be exclusive or exhaustive. For example, "percent acceptable growth" is rarely used in practice, and the metrics listed under growth prediction for the categorical model are listed in parentheses to emphasize that they are largely theoretical. Part of the challenge of validation is that new metrics can be created from the same raw data on the whims of any analyst at any time. These metrics can imply or inform different uses, adding new columns and new rows to Table 1 far faster than validation efforts can follow. Nonetheless, the metrics listed in Table 1 are good examples to illustrate the thesis of this chapter, that growth metrics differ, and that these differences interact with criteria such that some metrics are good for some things, and other metrics are good for others. In the next subsections, I briefly review each growth model and some of the metrics that they support.

The Gain-Based Model

The first and arguably most intuitive growth model is the one seemingly defined in the Race to the Top guidance as, "the change in achievement data for an individual student between two or more points in time" (U.S. Department of Education, 2009, p. 59742). The aliases listed in Table 1 are generally synonymous with this idea, that there is a trajectory that each student has over time that can be described as a slope or a gain. The statistical foundation is the simple difference between a current year or otherwise recent score and a past year or otherwise past score. Extensions are straightforward and include estimating trajectories over more than two points in time or allowing for nonlinear trajectories, in the tradition of longitudinal data analysis (e.g., Singer & Willett, 2003). However, as a matter of policy, models are generally kept fairly simple.

Gain-based models can support both growth description and growth prediction, for both students and schools. A commonsense metric for student-level growth is the gain score, the simple mathematical difference between current and past scores. Gains can be compared to "acceptable gains" by some standard setting procedure, and a school level metric could be the percentage of acceptable gains. Although gain scores are intuitive, they become problematic in that they rely on vertical scaling decisions, whereby expected gains may differ in magnitude across grades. Although average gains may be an accurate representation of the amount of learning in each grade on an absolute scale, this may be more attributable to the typical developmental trajectories of children than to schools. Straight comparisons of gains on developmental vertical scales are thus inappropriate for comparing the amount of growth for which schools may be responsible.

Typically, averaging across grades is only done after setting different standards within grades, as in a "percentage of acceptable gains" metric. However, vertical scales may also be problematic even with these adjustments, as higher scoring students may differ in expected gains over lower scoring students within any grade. This motivates many of the subsequent models, particularly those that use conditional status. In spite of these issues, gain-based models are linked closely to intuitive definitions of growth, and, as I argue later in this chapter, their transparency is an asset.

An alternative approach to setting standards for acceptable growth is to make predictions about the future and rely on standards set at that future time point. The trajectory model is a gain-based approach to making predictions about growth to a future time point. The most straightforward approach is to assume that the past gain extends into the future at an identical rate. If a student scored a 10 last year and a 15 this year, then the trajectory model suggests that the student will score a 20 the next year and a 25 the year following. I introduce some informal notation here, where X_g refers to a student score X at some grade g. This allows representation of a student gain score from grade 6 to grade 7 as: $X_7 - X_6$. In the manner of growth prediction, if one is interested in the predicted score at grade 8, the trajectory model estimate of this future status follows:

$$\hat{X}_8^{traj} = X_7 + (X_7 - X_6) = -X_6 + 2X_7. \qquad (1)$$

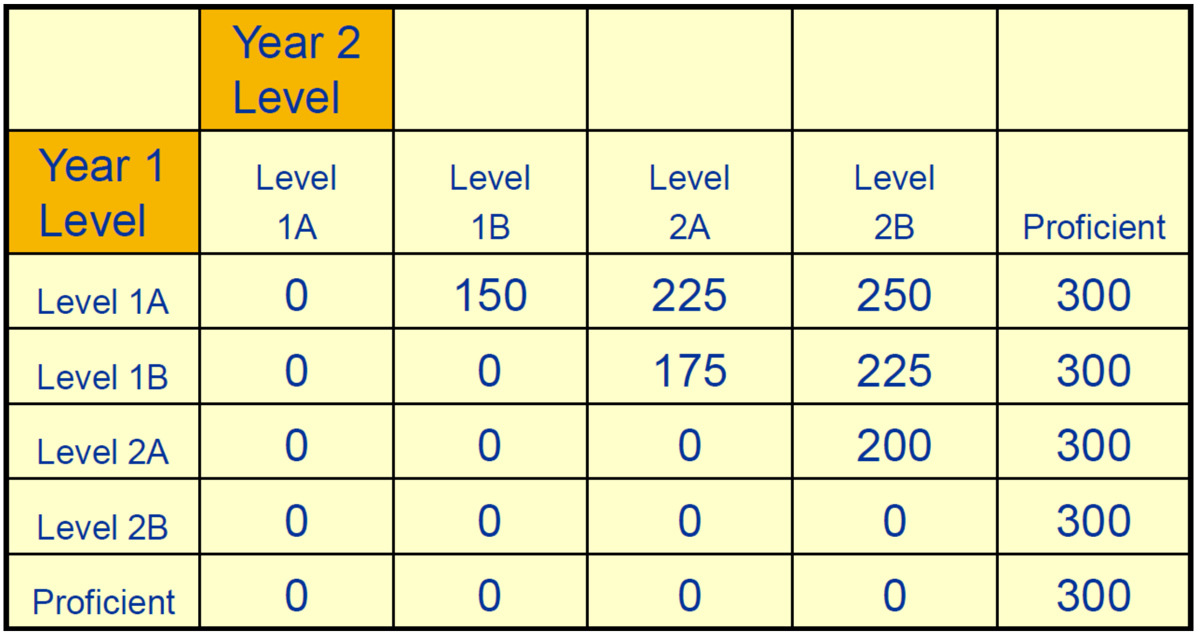
To determine whether this past growth is adequate, we may compare this student's predicted future score to some benchmark cut score, X_{8cut}. If $\hat{X}_8^{traj} \geq X_{8cut}$, then we may say that this student is "on track." A natural accounting of the school level growth is the simple percentage of students who are on track. An alternative approach, not listed in Table 1, is the average predicted future score. This contains no more information than the average gain score and suffers from the same dangers of vertical scaling as those mentioned earlier.

It is also worth noting that this expression of acceptable growth, "if the predicted score meets or exceeds the future standard," is equivalent to an alternative, seemingly different expression, "if the growth exceeds that needed to reach the standard." For this latter criterion to be met, the difference between grade 6 and the cut score at grade 8, or $X_{8cut} - X_6$, must be halved by grade 7 to be on track, $X_7 \geq \frac{X_{8cut} - X_6}{2} + X_6$. This is an entirely equivalent expression to $\hat{X}_8^{traj} \geq X_{8cut}$.

The Categorical Model

The categorical model, also known as the value table or the transition matrix model, divides each within-grade score scale into a smaller number of ordered categories (Hill, et al., 2006). Table 2 shows Delaware's value table for the 2009-2010 academic year (Delaware Department of Education, 2010). A student who scores in Level 1A in Year 1 but Level 2A in Year 2 receives a growth score of 225, as shown, and the average across students in the school represents the school-level score. The model relies more than others on the selection of cut scores, where transitions between categories function as student growth data. Logically, the cut scores between Level 1B and Level 2A must have some basis for equivalence.

Table 2. An example of a categorical model from Delaware's 2009-2010 school year.



	Year 2 Level				
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf

The categorical model is flexible in the sense that values for particular transitions between categories can be adjusted to user specifications. As Hill, et al. (2006), demonstrate, careful selection of values for particular transitions can result in a pre-growth-era status model, where only proficiency is counted, or something that seems more gain-based, where the gain is quantified as the number of levels that are gained or lost. The cost of this flexibility is the loss of information that comes with categorization, where the model cannot distinguish between the very highest and the very lowest scores in any given category. Although this may seem inappropriate for comparing growth for individual students, at the aggregate level, the errors due to coarse categorization are diminished, particularly as the numbers of categories increases. However, at a certain number of categories, judgments that support differing values become more difficult to distinguish and justify, and the model becomes likely to reduce to something similar to a gain-based model, with a number of categories approaching the number of score points.

The categorical model technically provides growth descriptions. However, the values that are selected for the categorical model may be motivated by inferences about whether a particular transition between categories is sufficient to warrant an "on track" designation for students making that transition. To the extent that these inferences inform the choice and interpretation of values, the function of a categorical model is one of growth prediction, as well as growth description. In the case of some growth models, like Iowa's model under its Growth Model Pilot (Hoffer, et al., 2011), this took the form of values like those in Table 2, except any nonzero value was simply a 1. This was based in part on the argument that a gain in categories established students as on track to proficient. In this way, the categorical model can support both growth descriptions and growth prediction. I evaluate models on the basis of predictive accuracy in the second half of this chapter.

The Student Growth Percentile Model

Betebenner (2009) introduced the Student Growth Percentile (SGP) metric as a normative approach to describing student growth. The SGP metric uses nonlinear quantile regression to support conditional status interpretations, where the current status of a student is referenced to expected percentiles given the score history of students. Although the name of the metric seems to indicate a percentile rank of growth scores as measured by gains, the statistical foundation is one of conditional status, where a student's current status is considered in light of expectations given past scores.

Castellano and Ho (2013b) review the SGP estimation procedure in detail. Fitting the statistical model can be time consuming for large datasets and uses an open-source R library (Betebenner, 2013). The SGP is calculated by first estimating 100 nonlinear quantile regression manifolds, for quantiles from .005 to .995, where the outcome variable is the "current" score and the predictor variables are all prior year scores. Castellano and Ho (2013b) demonstrate that this is practically similar to a straightforward linear regression model of current year scores on past year scores, where the SGP corollary is the percentile rank of residuals. In the case of SGPs, the nonlinear manifolds may cross, particularly at the extreme score ranges. The SGP package implements an "uncrossing" procedure to prevent nonmonotonicity, where higher scores might receive lower SGPs even conditional on past scores. After uncrossing, any student with a observed score that is located between, for example, the .325 and the .335 quantile manifolds receives an SGP of 33. The school-level SGP metric used most often in practice is the median SGP, which Betebenner (2008) has argued for on the basis of the ordinal nature of percentile ranks.

The SGP package also contains an option for growth prediction, in the form of Student Growth Projections (Betebenner, 2013). These are an intriguing hybrid of a trajectory model and a regression model. The projections are often displayed in a fan-shape spreading out from a student's current status (Betebenner. 2009), where higher portions of the fan correspond to the predicted score if a student earned a high SGP, and lower portions of the fan correspond to the predicted score if a student earned a low SGP. The scores that support these fan-shape displays are estimated from a previous or older cohort that has data relevant to the grade over which a prediction is made. For example, the data that supports a prediction for a 6th grader to her 7th grade year could arise from a previous cohort of 6th graders who now have 7th grade scores. Without these "reference cohorts," empirical predictions cannot be estimated.

Although Student Growth Projections support visual displays, they are also used to make specific predictions in practice (Betebenner, 2013). In order to be "on track," students are assumed to maintain their current SGP over time. This is an explicit prediction of future status and growth. If a student's future score exceeds some cutoff such as "proficiency," then the student is determined to be on track. Equivalently, if the student's SGP exceeds the minimum SGP that must be maintained to reach the future cutoff, then the student is on track. The logic of the equivalence of these two statements parallels the analogous equivalence demonstrated in the section about trajectory models. Importantly, as I will demonstrate, the assumption that students maintain their current SGP over time, rather than maintain a more neutral SGP of 50, is an appeal to intuition more than statistics.

Student-level growth prediction can take on many reporting forms, from the fan-shaped graph mentioned earlier to a simple dichotomous judgment about whether a student is on track. However, the direct implication of the student-level Student Growth Projection metric, free from the vagaries of standard setting in future grades, is captured by the actual predicted score in the future, assuming the current SGP is maintained. A school-level metric could be constructed by averaging predicted scores, although differing score scales across grades would likely make this problematic. A simpler school-level metric is the percentage of students who are on track: an average of the dichotomous student-level judgments. Again, many alternative aggregation schemes exist in the expanding universe of accountability models, such as the "adequate growth" designation in Colorado (Colorado Department of Education, 2009).

The Projection Model

The projection model uses more conventional regression techniques to describe and predict growth. The "projection" descriptor most often refers to the purpose of growth prediction, and Castellano and Ho (2013a) describe the model as serving this purpose in their guide. However, the regression-based statistical foundation lends itself well to growth description, also, particularly in the form of residuals or, as they are occasionally described (sometimes with due criticism, e.g., Rogosa, 1995), "residual gain scores." The student-level score is the simple difference between a student's observed score and her expected score given past scores. For a seventh grader with one prior-grade score from grade 6, this can be expressed simply as $e_7 = X_7 - \hat{X}_7$, where $\hat{X}_7 = b_0 + b_6 X_6$. The regression parameter estimates, b_0 and b_6, can be estimated by simple ordinary least squares, although a variety of alternative models and estimation procedures are available to suit the particular features of the data.

A practical feature of the projection model is that it does not require a vertical scale or any argument for a common scale across grades. As an expression of a deviation from an empirical expectation, the scale of the residual is the scale of the outcome variable, that is, the scale of the current-grade score. An alternative approach proposed by Castellano and Ho (2013b) involves taking the percentile rank of a student's residual in the distribution of all residuals. As they note, this Percentile Rank of Residuals (PRR) metric is nearly indistinguishable from SGPs in many real data scenarios.

For a school-level metric, a school-level average of residuals would be one approach, although this would be problematic when scales are not comparable across grades. This problem could be addressed in part by standardizing current-grade scores prior to regression, although an average across grades would nonetheless make the implicit assumption that standard deviation units are substantively equal across grades. Multilevel models are another possibility, where the school level metric could be a fixed or random intercept, although cross-grade comparisons remain complicated.

A simpler approach to a school-level regression-based metric involves taking a percentage of acceptable residuals, in the same way that a percentage of acceptable gains can be calculated for gain-based models. This requires cut scores articulated across grades, and, arguably, the assumption that cut scores are well articulated across grades is nearly as unrealistic as the argument that scales are comparable across grades. However, an exhaustive critique of all possible metrics is less the purpose of this chapter than acknowledging the proliferating number of metrics and the contrasting criteria on which they may be evaluated.

The projection model is particularly useful and is arguably even designed for optimizing prediction. For this purpose, as with Student Growth Projections, a reference cohort is needed that either uses a past cohort or an older cohort with the relevant grade-level data. For the purpose of predicting a future grade 8 score, for example, the projection model estimate takes the following form,

$$\hat{X}_8^{proj} = b_0^* + b_6^* X_6 + b_7^* X_7 \qquad (2)$$

Here, the b statistics have asterisks to denote that they are estimated not from the current cohort, which likely does not yet have grade 8 scores from which to estimate these parameters. Instead, these parameters are estimated from data from a past or older reference cohort. In contrast with Equation 1, which has no constant and fixed coefficients of -1 and 2, respectively, the coefficients for projection models are generally positive, with coefficients of larger magnitudes linked to proximal grades where partial correlations are higher. The weights of the projection model are empirically derived, whereas the trajectory model represents more of an aspirational, theoretically driven prediction.

The individual-level growth prediction is $\hat{X}^{proj}$. As with previous metrics, this predicted score can be compared with a cut score, $\hat{X}_{cut}$, and an "on track" designation may be assigned when $\hat{X}^{proj} \geq \hat{X}_{cut}$. A possible school-level metric is then the percentage of students who are on track. Importantly, the growth description and growth prediction functions are more fundamentally distinct here than in the gain-based, categorical, and SGP models. Although residual metrics and predicted scores use the same underlying regression machinery, the scores for the residual metric are clearly residual-based, focusing not on the predicted current score but the discrepancy between the observed score and this expectation. In contrast, the growth prediction machinery focuses solely on the expected future score, and in fact has no observed future score from which to frame departures from this expectation.

The distinction between growth description and growth prediction is less clear under the SGP approach, where the two concepts are linked conceptually. By assuming SGPs are maintained into the future, the Student Growth Projections continue to be driven by residuals. As I will demonstrate in the remainder of this chapter, this intuitively appealing consistency comes at the cost of predictive accuracy.

Contrasting Criteria for Evaluating Growth Models

In the second half of this paper, I contrast criteria for evaluating growth models. I focus on three in particular, predictive accuracy, transparency, and incentives. The first criterion, predictive accuracy, is largely applicable to growth metrics from the right side of Table 1. For students, this includes the trajectory model, student growth projections, and the prediction-oriented projection model. Predictive accuracy is a criterion that lends itself well to quantification. For continuous outcomes, a straightforward criterion is the root mean square deviation, $RMSD$, for example, for a grade 8 score,

$$RMSD = \sqrt{\boldsymbol{E}\left(\left(X_8 - \hat{X}_8\right)^2\right)} \tag{3}$$

This is estimable as the square root of the average squared deviation between observed and predicted scores. The RMSD is interpretable as the expected magnitude of the discrepancy between observed and predicted future scores. However, in the context of school evaluation and accountability, I argue that it would be shortsighted to select a model based on predictive accuracy alone. As a point of reference, it is trivial to assert that a dataset that allowed for perfect prediction would be not only unrealistic but undesirable in the context of education. If knowing past scores allowed for extremely precise predictions of distal future scores, that would be a damning testament either to the relevance of all intervening activity between present and future or the relevance of the outcome measure itself. At best, one would wish for a data structure that allowed for precise prediction in the absence of a treatment but enabled policy models that would degrade that prediction over time, as, presumably, the effect of the policy model had some impact.

Arguably, then, the point is not to have a model that makes predictions that end up being true, but a model that provides accurate information about what will happen if no intervention is taken, and a model that encourages interventions that result in this prediction being, ideally, biased in the negative direction. Nonetheless, it is desirable, all else equal, to have a model that affords accurate predictions, as this will best inform users about how to render model predictions inaccurate in the negative direction. From this perspective, I introduce the $RMSD$ as a criterion, but not the sole criterion from which to evaluate models, and I evaluate models along this criterion in the next section.

The second criterion, transparency, is far less amenable to quantification. I hesitate to label any of these models as inherently transparent or lacking in transparency, as clear explanations and reporting to key constituents can increase transparency. One might imagine a transparency criterion that asks target constituents whether they can replicate the model results themselves. Another facet of transparency is the ability of users to explain the growth metric to others. As students, parents, teachers, and administrators are increasingly invested in and affected by assessment results, their ability to replicate and interpret the results, or feel that they could, is essential. Whereas a statistical model that leads to a research finding may be a black box to the vast majority of the public, an accountability model will be less effective if it is perceived as opaque.

The third criterion concerns the incentive structures that the model and metric support. One might evaluate the incentives of a model by asking, is the model likely to lead to desired responses by constituents? This said, incentives are less a property of a particular model as much as a property of a particular metric. As noted in the introduction, the increasingly complex and layered design of accountability metrics can alter the incentive structure beyond what any underlying model may initially establish. For simplicity, however, I will address the incentives of particular models and metrics, in general, to establish lines of contrast between models of interest, acknowledging that alternative downstream manipulations of metrics may alter incentive structures. I will also reflect on interactions between the second and third criteria, where transparency generally works to enhance the impact of the underlying incentive structure, whether those incentives are desired or not.

Predictive Accuracy

In their final report of the GMPP, Hoffer, et al. (2011) used real data analyses to compare the predictive accuracy of three of the four models in Table 1, excepting SGPs. They found consistent rankings, where the projection model had the highest predictive accuracy, and the trajectory and transition models were a somewhat distant second and third, respectively. In this section, I provide some theoretical results that frame the findings and anticipate magnitudes of differences more generally.

First, for convenience, I consider a generic correlation matrix, $\mathbf{R}$, for cross-grade scores, with a common between-grade correlation of r. The matrix requires at least three grades, a current grade, a past grade, and a future grade to which scores are predicted. Note that this functions as the data from a reference cohort, as no current students have future scores from which to estimate projection models.

$$\mathbf{R} = \begin{bmatrix} 1 & \cdots & r \\ \vdots & \ddots & \vdots \\ r & \cdots & 1 \end{bmatrix}$$

This is a caricature of a real correlation or covariance matrix, and there are three practical concerns about its generalizability. First, the correlation matrix does not reflect the general and unsurprising finding that correlations between proximal grades are higher than correlations between distal grades. Second, the correlation matrix does not reflect the general finding that correlations between higher grades tend to be higher than those between lower grades (see Castellano & Ho, 2013b, for examples). Third, by using a correlation matrix instead of a covariance matrix, the variance of scores within grades is constrained to be equal. Although this is a fairly common feature of many across-grade scales, many developmental score scales display dramatic increases in variability toward higher grade levels. This does not affect the practical results for regression-based models but has implications for the accuracy of the trajectory model. Nonetheless, the benefits of this artificial representation is a clean demonstration of the relative predictive accuracy of models.

For example, for the projection model, we can use r to refer to the correlation, and k to the number of total scores that students have from which they may make predictions. The correlation matrix for grades 6 (past grade), 7 (current grade), and 8 (projected grade) is the matrix $\mathbf{R}$ with dimensions 3×3, and here $k = 2$ available scores. For a current grade $(g-1)$ to some future grade, g, one year into the future, the estimated projection equation, following Equation 2, is, $X_g^{proj} = b_0 + b_{g-1}X_{g-1} + \cdots + b_{g-k}X_{g-k}$. No asterisks are necessary here, assuming that the correlation matrix for the target sample matches that from the reference sample. From basic regression principles, we can derive the $RMSD$ for the projection model as follows:

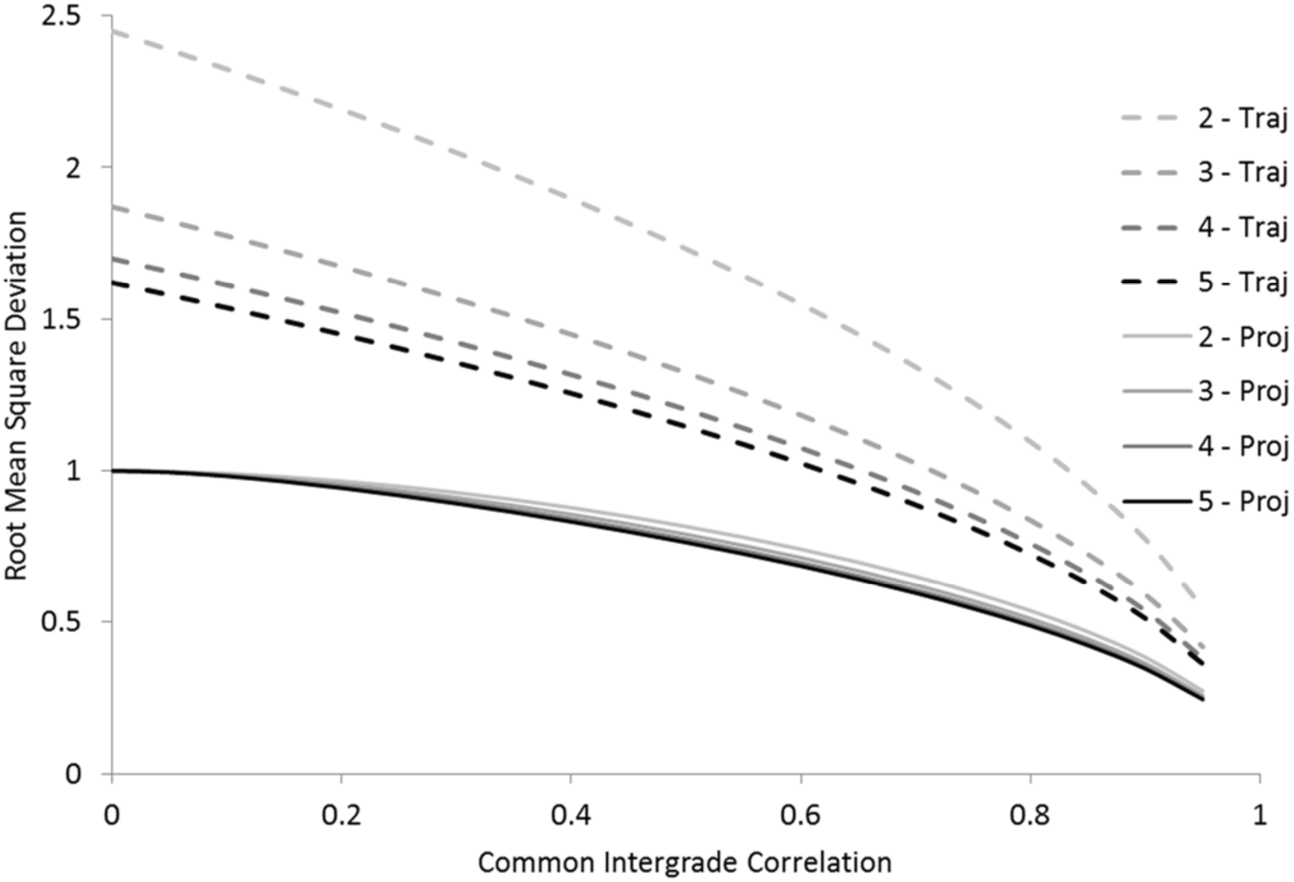
$$RMSD_{proj} = \sqrt{1 - \frac{kr^2}{1 + (k-1)r}}$$

Similarly, we can derive the $RMSD$ for the trajectory model, under the assumption that an "average gain" model is used, where the difference score is calculated from by subtracting the earliest grade score from the current grade score. This is equal to the average of consecutive gains from year to year. For example, if a student's score trajectory when $k = 3$ is 5, 10, 25, then averaging the two gain scores, $(5 + 15)/2 = 10$, is equivalent to taking the current grade score, subtracting the earliest grade score, and averaging by the number of years, $(25 - 5)/2 = 10$. The predicted score, following Equation 1, is thus, $\hat{X}_g^{traj} = X_{g-1} + \frac{X_{g-1} - X_{g-k}}{k-1}$. Under these assumptions, and using the same correlation matrix, $\mathbf{R}$, we can derive the RMSD for the trajectory model as follows:

$$RMSD_{traj} = \sqrt{2(1-r)\left(1 + \frac{k}{(k-1)^2}\right)}$$

Figure 1 shows the root mean square deviations over nonnegative correlations for $k = 2 \dots 5$ available scores. It is clear that the predictions of the trajectory model are dramatically worse than those of the projection model. For illustrative magnitudes of grade-to-grade correlations, from 0.6 to 0.8 (see Castellano and Ho, 2013b, for examples), the RMSDs for projection models are from 0.5 to 0.75 standard deviation units. The trajectory model RMSDs are 1.5 to 2.0 times as large, and this is the same factor by which confidence intervals would be larger. In absolute magnitude, the trajectory model RMSDs range from around 0.75 to 1.5 standard deviations in magnitude. This is a considerable amount of additional predictive error.

Figure 1. Theoretical root mean square deviations for prediction of a future grade score one year in the future. Results shown for projection and trajectory models, by common intergrade correlations and the number of available years of data for prediction.



For categorical models, the predictive accuracy depends upon the number and location of cut scores as well as the values in the value table. Values can be chosen that make the model more like a trajectory model, or values may be chosen that make the model function more like a regression model (Hill, et al., 2006). If values are chosen to match any particular model, however, one can be assured that the predictive accuracy will be lower given the loss of information inherent in categorization. Hoffer, et al. (2011), show that, in terms of correct classification rates, categorical models designed to match trajectory models have lower correct classification rates, sometimes by one or two percentage points but also by more than 10 percentage points depending on the subject and grade tested.

To estimate RMSDs for SGPs, consider the percentile rank of residual (PRR) analog to SGPs introduced by Castellano and Ho (2013b). They show very high correlations with SGPs as well as higher RMSDs when regression assumptions are met. In an ordinary regression context, we may extend the PRR framework to mimic the Student Growth Projection framework. As described in the previous section, Student Growth Projections are akin to making a prediction and then adding back the residual that corresponds to the SGP. To understand the implications, we first acknowledge that residuals from a current regression, say X_7 on X_6,and residuals from a future regression, say, X_8 on $X_7$and X_6, are uncorrelated by design.

Next, by adding the residual in grade 8 that corresponds to the percentile rank of residuals in grade 7, we add back the same amount of conditional variance in Grade 8 as the Grade 8 error variance itself. This effectively doubles the prediction error, and we obtain the result,

$$RMSD_{prr} = \sqrt{2}RMSD_{proj}.$$

When there is prediction one year into the future, the $RMSD$ of student growth projections is at least $\sqrt{2} \approx 1.414$ times that of projection models if regression assumptions are met. In practice, with prediction one year into the future, the predictive accuracy could be slightly worse if regression assumptions are met, as the quantile regression splines will result in overfitting, or the accuracy could be slightly better if regression assumptions are not met and the SGP model succeeds at capturing the shape of the multivariate population distribution. If predictions are extended farther into the future, as they are in practice, the predictive accuracy of projections will decline considerably, as residuals are layered from one year to the next. In short, assuming a student retains residual "momentum" is unrealistic one year into the future but is even less realistic across subsequent years. The predictive accuracy of student growth projections thus predictably varies across students with different SGPs and declines for students with noncentral SGPs (closer to 1 and 99) and for projections multiple years into the future.

Transparency

Transparency is a slippery criterion that is not inherent to any metric but relies on the reporting of the metric and clarity of the explanations of the appropriate uses that the metric supports. Earlier in this section, I observed that transparency is difficult to quantify, and I imagined hypothetical approaches to operationalize transparency such as asking constituents whether they can replicate model results or teach others its tenets. I also argued that the impression that people have about a model may be more important than whether or not they actually understand how the model works. The question may arguably be about whether the actions that users take are defensible, not whether they can explain the precise chain of reasoning that leads to this action. I begin this discussion of transparency with the metric that highlights this contradiction, between transparency as a feeling about a metric and transparency as an ability to understand a metric: SGPs.

I alluded to an initial argument against the transparency of SGPs in an earlier section. The metric appears to suggest a percentile rank of scores, particularly those that might be operationalized by a gain-based difference score or a slope estimated through a score trajectory over time. This percentile rank of gain scores is not what the metric represents. It instead provides a location in the conditional distribution of current scores given past scores. This is conditional status, as reflected in Table 1. A second charge against the SGP metric may be levied on the basis of the complex statistical procedures that support estimation. A routine that estimates 100 nonlinear quantile regression manifolds through a multidimensional surface and then "uncrosses" crossing manifolds is difficult even for many trained statisticians to follow.

On the other hand, these arguments have clear rebuttals. Arguably, a user who misinterprets an SGP as the percentile rank of a gain score may also confound that interpretation with the interpretation that is actually supported by the metric: Is the student performing better than expected given past scores? The percentile rank of a gain score can be deeply problematic if absolute gain scores are not of interest. Attribution of high percentile ranks to student, teacher, or school effort may be confounded by scaling issues even more than conditional status metrics are. By this argument, users who make inferences about absolute growth but desire information about expectation-referenced growth are being given the information they desire even if they do not know that they desire it.

The complexity of the SGP model is also offset by a remarkable degree of accessibility via the statistical program, R, which is free and whose libraries are open-source (Betebenner, 2013). This allows the procedures to be used, evaluated, and expanded by anyone who has sufficient technical proficiency. In addition, the SGP model comes packaged with a set of striking visual displays of information, some of which are included in Betebenner's 2009 article, and others of which are easily accessible in online score reporting tools such as http://www.schoolview.org/. Although a user may not be able to articulate precisely how an individual or aggregate-level SGP statistic should be interpreted or what its implications may be in the face of error, the smooth reporting interface and complete accessibility of its procedures represents a form of transparency that rivals and exceeds many other statistical metrics that function in accountability models.

The categorical model rates highly on the transparency criterion. By providing users clear tables with explicit values, Hill, et al. (2006), demonstrate that interpretations about what sorts of improvements are needed, and for which students, are straightforward. The notable threat to transparency is the defensibility of the standard setting process by which cut scores are set and the setting of values associated with particular transitions. If cut scores are not articulated to carry the same meaning across grades and over time, interpretations about the kinds of progress needed will be flawed. Castellano and Ho (2013a) describe the process of setting cut scores and values in a categorical model as that of defining an "implicit vertical scale," where the cut scores and values interact to effectively assign a weight to transitions along a particular region of a vertical scale. Even if no vertical scale actually exists, the weights and cut scores create an effective scaling that has implications similar to that of a vertical scale.

The trajectory model has a similar level of transparency to that of the categorical model, although its transparency depends in part on a clearly anchored or otherwise well-defined vertical scale. It operationalizes growth in an arguably intuitive fashion, where growth is measured along some vertical scale, and time is on the horizontal axis. In contrast, the projection model, which I argue is less intuitive, cannot be displayed graphically with time on the horizontal axis without some constraining assumptions. In a projection model, time is only represented to the extent that correlations between proximal grades are higher than correlations between distal grades. Although the metaphor of "projection" in a growth model suggests and extension over time, the estimation is based on regression and thereby conditional status.

Both the trajectory and the projection model can have their transparency enhanced by thoughtful displays of information. For the trajectory model, the natural representation is one of scores over time. For a projection model, the natural representation is an equation, with estimated weights, where users are able to plug in scores to either get an expected current score. This expected score may be compared to a current score, where the difference is a residual gain score. Or the expected score may be derived from an equation that was estimated on reference data, for a prediction of future status.

Comparing these past two sections, it is clear that maximizing predictive accuracy, which can be done with projection models for future status, may not be the path to maximizing transparency, whereas a transition model or a trajectory model may be more transparent. In the next section, I discuss a third criterion concerning the incentives that are set by growth models, for students or those associated with their aggregated scores.

Incentives

Responses to incentives depend in part upon transparency. If a score report does not communicate actionable differences among students or schools, users of model results will not know or care to respond to model incentives. Assuming adequate transparency, however, the obvious incentive supported by growth models is to maximize growth, however that growth is operationalized. This often manifests in a presumably desired response, where higher scores are achieved in current years, over and above past scores. However, there are alternative strategies for maximizing growth metrics that are unlikely to be desired responses.

As a general example, the trajectory model in Equation 1 is maximized when either current grade scores are high or when previous grade scores are low. A cynical approach to maximizing gain scores involves artificially deflating initial scores for the first grade that is tested, an approach that I refer to as "sandbagging," after a similar term in sports, from golf handicapping to concussion tests, where early performance is artificially deflated to make subsequent performance results seem high by comparison. As Table 2 makes clear, sandbagging is possible for categorical models, as well, and can be visualized as artificially moving up in any particular column, acquiring lower Year 1 levels for constant Year 2 performance, and earning higher scores as a result.

The extent to which sandbagging can influence subsequent gain scores varies across metrics. For gain-based models, attempting to zero out initial test scores dramatically affects gain scores, thus sandbagging is particularly useful. In contrast, sandbagging for categorical models is generally more muted by simple virtue of the fact that there is no differentiation in the lowest scoring category. The SGP metric is susceptible to more moderate sandbagging. Early in a student's growth trajectory, an initial low score can lower expectations to a degree approaching that of a gain-based model. However, as SGPs pool expectations over multiple years of testing, a low initial score will get muted by subsequent higher scores that will raise expectations and conditional quantiles. Given the close relationship between residual gain scores and SGPs, their susceptibility to sandbagging is similar.

As I have emphasized, incentives are a function of a particular metric rather than a particular model. When multiple overlapping policies incentivize multiple metrics, incentives for teachers may conflict with incentives for students or schools. For example, in school-level Adequate Yearly Progress calculations, students in their first grade in a school are not eligible for growth calculations, as they have no prior-year scores. Although sandbagging would increase their subsequent gain scores, this cynical action would end up classifying the student as nonproficient, which would count against the school in the current year. The ideal strategy for maximizing future gain scores while avoiding nonproficiency in the current year is to sandbag performance to the minimal score necessary for achieving proficiency, but no lower.

At the teacher level, incentives may be similarly specific. For a residual gain or median SGP metric, teachers are incentivized to have students with sandbagged scores from the prior year, regardless of proficiency considerations, as these status classifications only have an impact on adequate yearly progress designations at the school level. Sandbagging may be a viable and cynical strategy across almost all growth metrics, but the precise incentives vary by degree across metrics, and the layering of multiple metrics leads to different incentives functioning at different levels of aggregation.

As far as incentives for particular growth metrics, the projection model is the most unique. Compared to gain-based models, where Equation 1, for example, includes a negative weight on the coefficient for prior grades, projection models generally have positive weights. In Equation 2, for example, the slope coefficients for the illustrative correlation matrix, $\mathbf{R}$, are all positive and identical, at:

$$b = \frac{r}{1 + (k-1)r}.$$

In this case, sandbagging is not effective, as any artificial deflation of performance will decrease the predicted score. Regression coefficients estimated from real-world correlation matrices will certainly have different regression weights and higher weights on proximal-grade variables, but weights on early variables will not be negative. As a result, the incentive structure for projection models is unique.

This is not to suggest that the incentive structure of projection models is always preferable. As Hoffer, et al. (2011) and Ho (2012) describe, projection models tend to create a kind of "inertial status" for consistently low-scoring and consistently high-scoring students. For these students, a small number of low or high scores results in prediction that is hard to influence with more recent grade data. As a consequence, bizarre incentive structures can result, where, for example, a high-scoring student can score a zero and still be predicted to be on track, or a low-scoring student finds it impossible to be considered on track, even if she scores a perfect score. From the perspective of the projection model, these cases are rare and unrealistic. However, if the purpose of the model is to incentivize growth for individual students, the projection model has clear shortcomings.

Discussion

Table 3 reviews the past three sections and paints a rough picture of the four models evaluated along the three criteria. As the table title notes and I have described, the evaluation of growth metrics hinges on small details, thus the evaluations in Table 3 are made about each respective model in its most typical form, as their results are most often reported. I fully concede that operationalization of a particular model may change these rough, relative judgments. However, it is useful to have a baseline for comparison of the models in their generic forms, and the past three sections have presented my arguments for these evaluations.

Table 3. A rough overview of growth models evaluated along selected criteria. .

Model	Statistical Foundation	Predictive Accuracy	Transparency	Distorted Incentives	
				Growth Description	Growth Prediction
Gain-Based	Difference: Current Score Minus Past Score	Medium	High	Sandbagging - High Risk	Sandbagging - High Risk
Categorical	Categorical. Changes in Categories.	Low	High	Sandbagging - Moderate Risk	Sandbagging - Moderate Risk
Student Growth Percentile	Quantile Regression. Conditional Status.	Medium - Variable	Debatable	Sandbagging - Lower Risk	Sandbagging - Lower Risk
Projection	Regression. Conditional Status.	High	Medium	Sandbagging - Lower Risk	Inertial Status

Note: Evaluations are not inherent to the models themselves and can be mediated by additional decisions

I have alluded to small decisions that matter and review three in particular here. First, sandbagging in all of its forms can be disincentivized by layering status models over growth models, although attention to the level of aggregation at which each is consequential is crucial. Second, the low predictive accuracy of categorical models can be increased by adjusting values assigned to transitions and increasing the number of categories. At a certain point, however, this may decrease transparency and effectively creates a coarsened projection model. Third, student growth projections may be altered to increase predictive accuracy by selecting a central SGP, such as 50, to carry forward, instead of assuming current

SGPs will be maintained. This effectively functions as a projection model, as well, and raises the question of whether a more suitable predictive modeling framework would be preferable over nonlinear quantile regression.

This chapter has contrasted three criteria—predictive accuracy, transparency, and incentives—and demonstrated that growth models and metrics share particular and differing strengths and weaknesses along these criteria. It is tempting to try to maximize all three criteria subject to some constraints, but, as the previous paragraph demonstrates, there are clear tradeoffs. Predictive accuracy is the most easily quantifiable criterion and is thus a compelling target. However, the incentive structures associated with projection models are, in my opinion, pernicious. Although accuracy of future predictions may be useful for targeting instructional resources, I do not find it defensible to use an accurate prediction of, for example, low future status as a reason to effectively disqualify a low-scoring student from earning an "on track" designation. If anything, there should be additional incentives to teach these students. This, of course, is precisely what gain-based models do, and this is what opens them up to the risk of sandbagging.

On the other hand, transparency is a criterion that may be increased without necessarily diminishing accuracy or skewing incentives. The SGP package is in many ways a transparent score reporting package, and this has made what is undeniably a complex statistical procedure seem straightforward and attractive. Projection models should be similarly transparent and include clearly specified equations. Trajectory models may also be accompanied by the visual trajectories of growth over time, using not only linear specifications but regression-based procedures that set realistic expectations in the light of vertical scales that may be tenuous. Certainly, any metric may be gamed, and transparency is likely to increase the likelihood of both intended and unintended responses. Presumably, however, this is preferable to a policy model whose impact is limited by its opacity.

We are at a point in history where state tests may come into increasing alignment under the so-called Common Core. However, there is very little indication that accountability models will experience the same alignment, and their complexity seems more likely to increase than decrease. As growth models continue to proliferate, it becomes all the more important to have clear specification of not only the statistical model but all of its related metrics, and all of their functions at all levels of aggregation. As I have argued here, all of the small details matter. For each metric, I hope that this chapter has sketched useful criteria and made some of the likely tradeoffs clear.

References

Betebenner, D. W. (2008). *A primer on student growth percentiles.* Retrieved from the Georgia Department of Education website: http://www.doe.k12.ga.us/

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

Betebenner, D. W. (2013). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories. [R package version 1.1–0.0].

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models.* Council of Chief State School Officers,

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics, 38*, 190-215.

Colorado Department of Education (CDE). (2009). *The Colorado growth model: Frequently asked questions*. Retrieved from http://www.schoolview.org/GMFAQ.asp

Delaware Department of Education. (2010). For the 2009-2010 school year: State accountability in Delaware. Retrieved from http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value-added models of student performance* (pp. 255-290). Maple Grove, MN: JAM Press.

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.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-65), Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

U.S. Department of Education (2005, November 21). Key policy letter signed by the Education Secretary. Retrieved from http://www2.ed.gov/policy/elsec/guid/secletter/051121.html

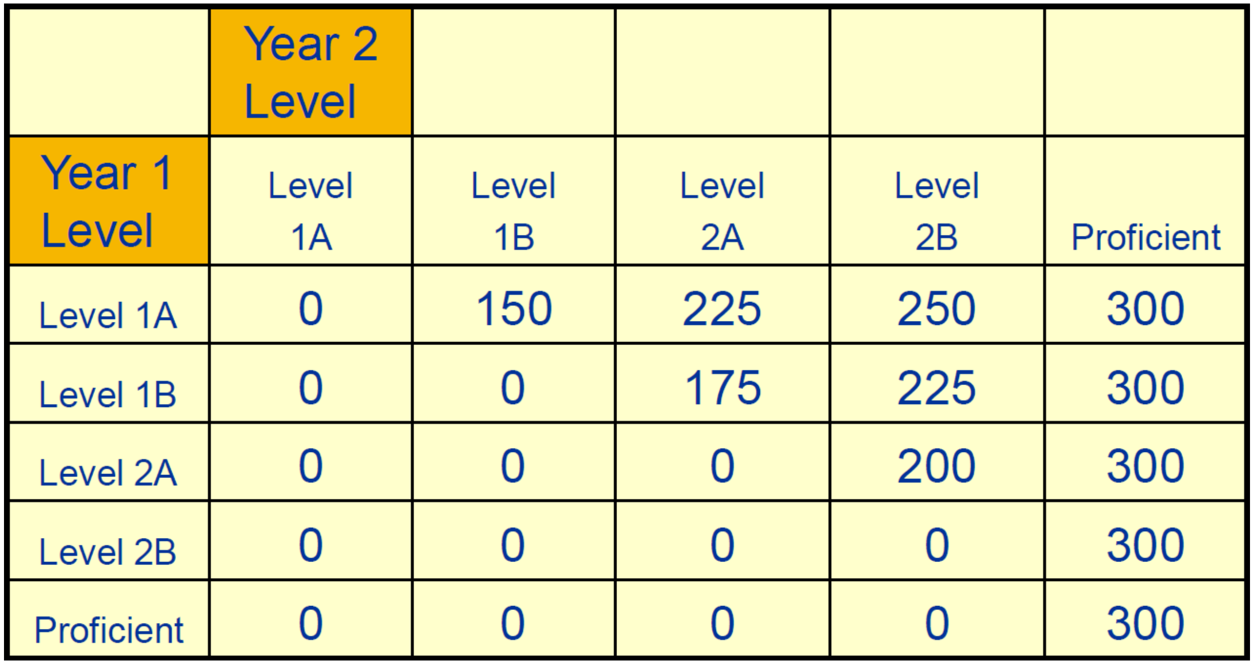
U.S. Department of Education (2006, January 25). Peer review guidance for the NCLB growth model pilot applications. Retrieved from http://www2.ed.gov/policy/elsec/guid/growthmodelguidance.doc

U.S. Department of Education (2009, November 18). Race to the top fund. *Federal Register*, *74*(221), 59688-59834. Retrieved from http://www.gpo.gov/fdsys/pkg/FR-2009-11-18/pdf/E9-27426.pdf

Table 1. Growth models, their statistical foundations, and the growth metrics they support.

Model	Aliases; Related Terms	Statistical Foundation	Growth Description		Growth Prediction	
			Student Level	School Level	Student Level	School Level
Gain-Based	Trajectory, Slope, Difference, Gain	Difference: Current Score Minus Past Score	Gain Score	Percentage of Acceptable Gains	Trajectory Model	Percentage On Track
Categorical	Value Table, Transition Matrix	Categorical. Changes in Categories.	Value of Category Change	Averaged Values of Category Changes	(Projected Category)	(Average Projected Category)
Student Growth Percentile	Colorado Model, Betebenner Model, Percentile Performance Index	Quantile Regression. Conditional Status.	Student Growth Percentile	Median Student Growth Percentile	Student Growth Projection	Percentage On Track
Projection	Regression Model, Residual Model, Multilevel Model, Hierarchical Model	Regression. Conditional Status.	Residual, "Residual Gain"	Average Residual	Projection Model	Percentage On Track

Table 2. An example of a categorical model from Delaware’s 2009-2010 school year.



	Year 2 Level				
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

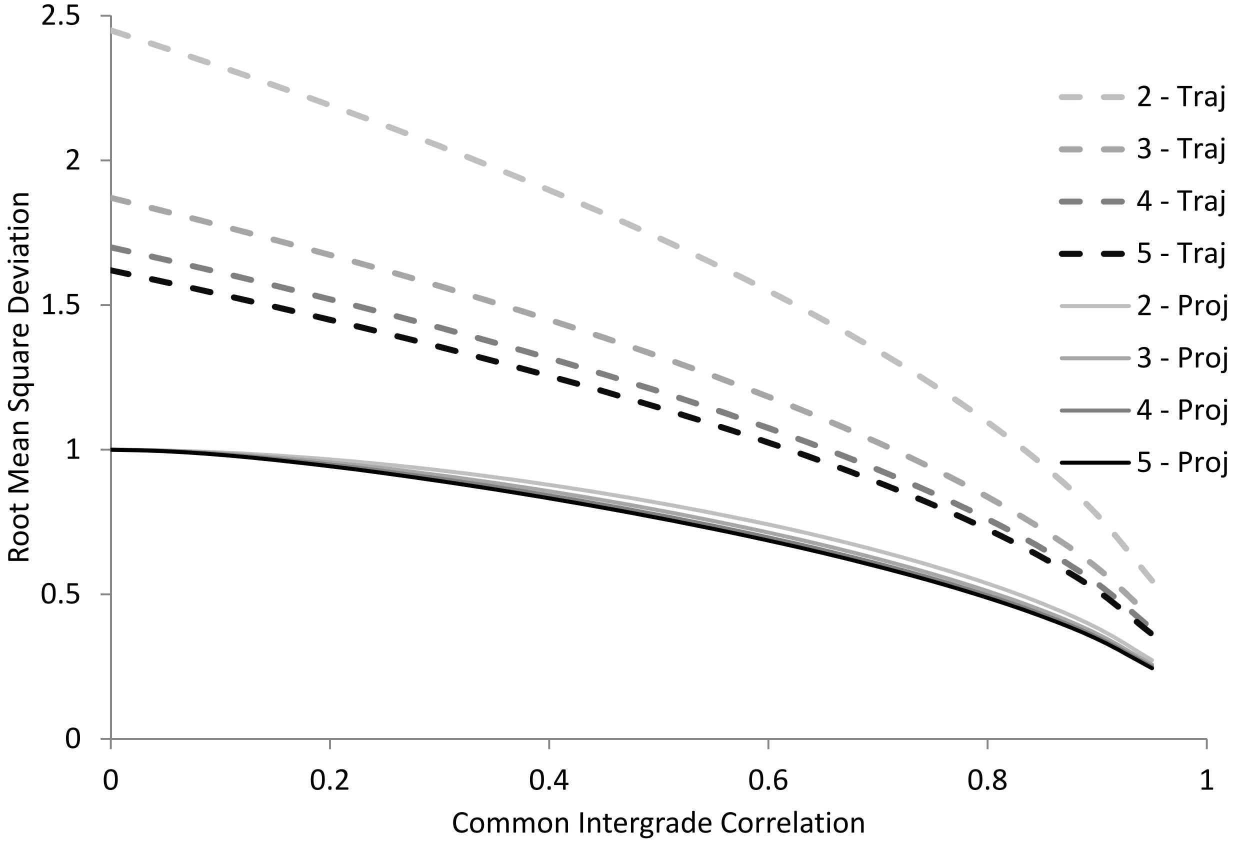
http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf

Table 3. A rough overview of growth models evaluated along selected criteria.

Model	Statistical Foundation	Predictive Accuracy	Transparency	Distorted Incentives	
				Growth Description	Growth Prediction
Gain-Based	Difference: Current Score Minus Past Score	Medium	High	Sandbagging - High	Sandbagging - High
Categorical	Categorical. Changes in Categories.	Low	High	Sandbagging - Low	Sandbagging - Low
Student Growth Percentile	Quantile Regression. Conditional Status.	Medium - Variable	Debatable	Sandbagging - Medium	Sandbagging - Medium
Projection	Regression. Conditional Status.	High	Medium	Sandbagging - Medium	Inertial Status

Note: Evaluations are not inherent to the models themselves and can be mediated by additional decisions.

Figure 1. Theoretical root mean square deviations for prediction of a future grade score one year in the future. Results shown for projection and trajectory models, by common intergrade correlations and the number of available years of data for prediction.


Root Mean Square Deviation
Common Intergrade Correlation
2 - Traj
3 - Traj
4 - Traj
5 - Traj
2 - Proj
3 - Proj
4 - Proj
5 - Proj