Evaluating the Use of Growth Prediction Models to Inform Instruction

Paper Presented at the National Council on Measurement in Education San Francisco, CA

Daniel Murphy

April 2013

ALWAYS LEARNING PEARSON

Abstract

This study examines the performance five growth prediction models (trajectory, transition table, projection, Student Growth Percentile, logistic regression) using data from two cohorts of students (elementary and middle school) in two subjects (reading and mathematics) with two proficiency cut scores (low and high rigor) on a vertically scaled summative assessment from a large U.S. state. This study evaluates the growth models for the purpose of supporting instruction by examining their accuracy when predicting declining or improving student performance. Results suggest that predictive value depends on target grade proficiency rates. Further, results suggest that models relying on assumptions of average growth provide better predictive quality than those relying on assumptions of constant growth. Tradeoffs among predictive value, type I, and type II error rates are inevitable and difficult to negotiate when predicting declining or improving performance with binary indicators. Logistic regression models may offer an attractive alternative to commonly used growth prediction methods.

Keywords: on-track indicators, growth prediction

Evaluating the Use of Growth Prediction Models to Inform Instruction

There is currently interest in measuring growth to inform instruction using the next generation of assessments. Both the Partnership for the Assessment of Readiness for College and Careers (2010) and SMARTER Balanced Assessment (Washington State, & SMARTER Balanced Assessment Consortium, 2010) Consortia have proposed the use of growth models to provide *actionable information* about students' progression to the Common Core State Standards' overarching goal of bringing all students to college and career readiness. Growth prediction models that are currently used for school accountability purposes have the potential to provide such actionable information.

Growth prediction models (Castellano & Ho, 2012) currently used to provide on-track designations in U.S. public school systems were developed "to determine whether measuring individual student growth over time would be another appropriate way to determine adequate yearly progress (AYP) under the Title I program" (U.S. Department of Education, 2006). When growth is defined in this manner, the main purpose of growth models is to give credit to schools whose students are not meeting current proficiency standards but are progressing such that they will be proficient in the future. By the same logic, growth models should be able to identify the opposite trend – students who are currently proficient but are progressing such that they will not be proficient in the future (Briggs, 2011).

A careful look at this definition of growth reveals two growth model components: one component is a measurement of the amount of progress an individual student has achieved, and the second is whether that progress indicates a path toward future proficiency. Each growth model component has the potential to provide actionable

information within the classroom. The amount of growth an individual student has achieved can provide useful diagnostic information to answer questions such as:

- Did the student make a year's worth of progress?
- Does the student's growth compare favorably with other students in the class, school, district, and state?
- Did the student grow as much this year as last year?

The second growth model component allows teachers to see how much growth a student is predicted to make. This "on-track" prediction can serve as an early-warning indicator to identify at-risk students and allow for more timely interventions if it is reliable. For example, teachers might want to design separate action plans for students who are:

- Currently proficient and predicted to remain proficient
- Currently proficient but predicted to be non-proficient in the near future
- Currently non-proficient and predicted to remain non-proficient
- Currently non-proficient but predicted to be proficient in the near future

It is important to note, however, that on-track indicators are imperfect predictors of future performance; therefore, growth model predictive quality is an important issue.

Perfect prediction is not attainable – or even desirable – in practice. If on-track predictions were perfect, then a student's future performance would be determined by current status alone, which would imply that instructional intervention is futile. At the other end of the spectrum, assigning instructional interventions using biased on-track predictions would imply mismatches between students and interventions, which would be inefficient. In short, growth model prediction errors are inevitable, but too many errors

invalidate the model. Therefore, the degree to which growth prediction models provide actionable information depends on their predictive value.

This study examines the predictive value of generic versions of four models that are commonly used to measure AYP, and a fifth model that is not currently used for that purpose. Receiver operating characteristic (ROC) analysis (Peterson & Birdsall, 1953; Swets, 1973, 1988, 1996), which is commonly used in the fields of medicine and psychology to analyze predictive quality, is used to examine predictive value. In particular, the predictive value that the growth models provide can be assessed using the ROC analysis tools of sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV), which are described in Figure 1.

	Non-Proficient in Target Grade	**Proficient in Target Grade**	**Predictive Value**
Off-track	True Negative (TN)	False Negative (FN)	$NPV = \frac{TN}{TN + FN}$
On-track	False Positive (FP)	True Positive (TP)	$PPV = \frac{TP}{FP + TP}$
Sensitivity/ Specificity	$Specificity = \frac{TN}{TN + FP}$	$Sensitivity = \frac{TP}{FN + TN}$	

Figure 1. ROC analysis tools: sensitivity, specificity, NPV, and PPV.

The PPV and NPV values are discussed within the paper in terms of odds ratios: $PPV\ odds = \frac{PPV}{1-PPV}$, and $NPV\ odds = \frac{NPV}{1-NPV}$. PPV odds of 5:1, for example, would indicate that students assigned an on-track indicator are five times more likely to be proficient than non-proficient in the target grade. On the other hand, PPV odds of 1:1 would indicate no predictive value, because the odds of proficiency in the target grade given an on-track prediction are no better than the odds of non-proficiency, putting the

predictive value on a par with a coin toss. PPV and NPV odds ratios are used with the ROC characteristics of sensitivity and specificity to investigate the following questions:

1. What predictive value do the growth models provide when predicting students will maintain proficiency?
2. What predictive value do the growth models provide when predicting students will decline in performance from proficiency to non-proficiency?
3. What predictive value do the growth models provide when predicting students will maintain non-proficiency?
4. What predictive value do the growth models provide when predicting students will improve in performance from non-proficiency to proficiency?

The predictive value of five models (trajectory, transition table, projection, Student Growth Percentile, logistic regression) is examined and compared with a counterfactual status model using test score data from two cohorts of students (elementary and middle school) in two subjects (reading and mathematics) with two proficiency cut scores (low and high rigor) on a vertically scaled statewide assessment from a large U.S. state.

Background

The existing research indicates that growth prediction model accuracy varies depending on the type of growth model used and the length of time between the prediction and the outcome (Hoffer et al., 2010; Weiss & May, 2012; Weiss, 2008). Growth prediction models rely on different assumptions about the nature of student growth to make predictions. Trajectory models, for example, assume that the amount of growth measured during one year for an individual student will continue at the same rate

during all future years. Projection models, on the other hand, assume that each student will progress at the average rate seen by the cohort as a whole.

Projection-based models have been found to be more accurate than trajectory-based models, and the nearer in time the predictions are to the outcome, the more accurate they are. Growth model types also can exhibit differing patterns of bias depending on the proficiency trends in the data (Weiss & May, 2012; Weiss, 2008). When evaluated for the accuracy with which it predicted non-proficient students were on-track to proficiency, Florida's trajectory model, which assumes constant growth, produced substantially more type I than type II errors. In practice, growth model bias toward type I errors could result in students not receiving necessary interventions.

By contrast, models that assume the cohort average rate of growth (e.g., projection model) though more accurate, were found to produce more type II than type I errors when applied to the same data. Growth model bias toward type II errors could potentially waste resources by assigning interventions to students who do not need them. Because perfect prediction is impossible, type I and type II errors are inevitable and unavoidable. Therefore, careful consideration should be given to the costs associated with each.

It has also been found that simply assuming all students will remain in the same proficiency category (i.e., a status model) can predict future proficiency as well as the best growth prediction models (Weiss & May, 2012; Weiss, 2008). This finding may not be surprising when analyzing prediction accuracy for the entire student population. Students who are already scoring well above a proficiency cut score have much better chances of being on-track to future proficiency than do students scoring near or below the

cut score. Likewise, students who are scoring well below a proficiency cut score have much better chances of being off-tack to future proficiency than students scoring near or above the proficiency cut score. In other words, growth models and a status model are equally likely to accurately predict that high scoring students are on-track and low scoring students are off-track. Differences between the models are most likely to be seen for those students who are close to the cut score and are progressing in a manner that suggests they will change proficiency levels in the future. Unlike previous studies that have investigated classification accuracy for the student population, this study will isolate these particular students to examine the growth models' predictive value.

Models

The growth prediction models examined in this study are generically specified so that, to the extent practicable, variability across model results can be interpreted as due to inherent model differences, rather than differences in model specification. It is important to note that specific models within each of the four general growth model types can vary in practice.

Status Model

A true status model does not make assumptions about future performance. For the sake of this study, to provide a baseline counterfactual against which the performance of the growth prediction models can be measured, the status model is used to make future predictions by assuming that students' proficiency level will remain constant across time. That is, currently proficient students are predicted to be proficient in the target grade, and currently non-proficient students are predicted to be non-proficient in the target grade.

Trajectory Model

The trajectory model uses a linear trajectory from a student's first-year score to a target cut score *N* years in the future. A student's second-year score is considered to be on-track if it meets the conditions of the following expression:

$$X_2 + N\left(X_2 - X_1\right) \geq T\,,$$

where X_1 is the first year score, X_2 is the second-year score, N is the difference in years between X_2 and the target cut score, and T is the target cut score. The trajectory model relies on an assumption that the rate of change between an individual student's first year score and second year score will remain constant in the future to make its on-track predictions.

Transition Table

Transition tables subdivide assessment scale scores into categories that are inferred to be learning stages. Student progress is then tracked across learning stages from one year to the next. Students are considered to be on-track if the stage change progression across two years is sufficient to indicate that continued progress will achieve the target cut score. The transition table is very similar to the trajectory model when the subdivided stages have approximately equal scale score ranges, as is the case in this study. It also relies on an assumption that the rate of change between the first year score and the second year score will remain constant in the future.

Projection Model

The projection model uses current and past performance to predict future performance. For the purpose of this study, the model will use data from the current

(grade *g*) and previous (grade *g* – 1) grade to predict performance in the target grade (grade *g* + *N*). Thus, the regression equation for the generic projection model is

$$\hat{X}_{g+N} = \hat{\boldsymbol{\beta}}_0^* + \hat{\boldsymbol{\beta}}_1^* X_g + \hat{\boldsymbol{\beta}}_2^* X_{g-1}.$$

In this equation, the regression coefficients, $\boldsymbol{\beta}_j^*$, are denoted with an "*" to signify the use of a reference cohort to generate the estimates. In particular, a 2010 reference cohort is used to generate the regression coefficient estimates that are then applied to the 2011 prediction cohort data. If the predicted score is equal to or greater than the target cut score, the student is classified as on-track. In contrast to the trajectory and transition table models, the projection model relies on an assumption that students will progress at the cohort's average rate of growth to make its predictions.

Student Growth Percentiles

The Student Growth Percentile (SGP) model (Betebenner, 2009) examined in this study estimates a conditional quantile function,

$$\mathrm{Q}\left(\tau \middle| X = x\right) = \mathbf{x}'\boldsymbol{\beta}(\tau)$$

where $Q(\tau)$ is the τ th quantile of random variable *Y* and $\boldsymbol{\beta}(\tau)$ is the set of regression coefficients $\{\beta(\tau): \tau \in (0,1)\}$. For the purpose of this study, the current-year student growth percentile estimate is conditional on test scores from the current and previous grades and estimated using the Transreg and Quantreg procedures of SAS statistical software version 9.2 (SAS Institute Inc, Cary, North Carolina) in a manner designed to mimic as closely as possible the SGP package that is implemented in the R statistical environment (R Development Core Team, 2008). The conditional quantile functions are parameterized as a linear combination of B-spline cubic basis functions (Colorado Department of Eduction, 2008; Wei & He, 2006). The Transreg procedure is used to

estimate seven cubic polynomial basis coefficients for each predictor variable, and then the Quantreg procedure is used to estimate regression coefficients for quantiles that range from .005 to .995.

As with the projection model, a reference cohort is used to generate regression coefficients which are then applied to the prediction cohort. It is important to note that if all students were assigned an on-track prediction based on assumed growth at the 50th percentile, then the SGP model would rely on the same assumption of average growth as the projection model, and it is likely that the on-track predictions of the two models would be very similar. However, in practice the SGP model makes on-track predictions based on an individual student's conditional growth percentile estimate rather than an assumption of cohort average growth. For example, a student estimated to have grown at the 30th percentile will receive an on-track prediction based on whether or not continued growth at the 30th percentile will lead to proficiency. The SGP model's reliance on an assumption of constant growth rate is therefore more similar to the trajectory and transition table models than the projection model.

Logistic Regression

The logistic regression model is a type of projection model that uses current and past performance to predict future performance on a binary (i.e., proficient/non-proficient) outcome and reports outcomes in terms of probabilities. The model in this study will use data from the current (grade g) and previous (grade $g - 1$) grades to predict performance level in the target grade (grade $g + N$). Thus, the regression equation for the generic logistic regression model is

$$logit(p_i)_{g+N} = \hat{\beta}_0^* + \hat{\beta}_1^* X_g + \hat{\beta}_2^* X_{g-1},$$

where $logit(p_i)$ is the natural log of the odds of proficiency in the target grade and the regression coefficients are as defined previously. To identify a model-based threshold cut score to indicate on-track requires two steps. First, an ROC curve is produced when estimating the logistic regression coefficients, and then Youden's Index (Bewick, Cheek, and Ball, 2004; Fluss, Faraggi, and Reiser, 2005) is used to select the optimal threshold value *c* from those produced for the ROC curve using the following expression.

$$J = max_c\{sensitivity(c) + specificity(c) - 1\}$$

ROC curves are discussed in more detail in the next section of the paper. Students with predicted probabilities of proficiency below *J* are classified as off-track, and those with predicted probabilities equal to or above *J* are classified as on-track.

Method

Data

The reported analyses come from a large U. S. state's summative mathematics and reading test scores for elementary and middle school student cohorts who tested in 2011. The grades and subjects were selected in part to hold the data constant across models in an attempt to reduce confounding effects due to variability across data sources. For the elementary school cohort, mathematics and reading test score data from grades 3 and 4 are used to predict grade 5 performances (i.e., $N = 1$); for the middle school cohort, mathematics and reading data from grades 5 and 6 are used to predict grade 8 performances (i.e., $N = 2$).

Table 1. Correlations, Variances, Means, Sample Sizes, and Proficiency Rates for the Elementary-School Cohorts

Grade	3	4	5	Variance	Mean	Low-Rigor Proficiency	High-Rigor Proficiency
Reference Cohort[a]					Sample Size = 266721		
3		.67	.64	10282	613	.92	.40
4	.69		.70	10223	655	.87	.31
5	.68	.74		7868	713	.88	.36
Variance	7639	9683	9467				
Mean	586	662	708				
Low-Rigor Proficiency	.87	.87	.88				
High-Rigor Proficiency	.33	.42	.44		Sample Size = 271958		
Prediction Cohort[a]					Sample Size = 274600		
3		.68	.65	9817	620	.88	.29
4	.70		.69	9584	658	.88	.29
5	.69	.74		8154	716	.89	.40
Variance	8611	8798	9428				
Mean	596	663	707				
Low-Rigor Proficiency	.90	.90	.88				
High-Rigor Proficiency	.39	.39	.44		Sample Size = 285649		

[a]The statistics for the reading cohort is above the diagonal and those for the mathematics cohort is below the diagonal.

The predictive accuracy of each growth model is examined across low- and high-rigor cut scores set within each test at the target grade. These cut scores are empirical cut scores from the statewide test. The low-rigor cut score corresponds to a relatively lenient "basic skills" standard. By contrast, the high-rigor cut score corresponds to a more difficult "advanced skills" standard. The correlations, means, variances, sample sizes, and proficiency rates for both the elementary and middle school reference and prediction cohorts are presented in Tables 1 and 2 respectively. Students with missing records are excluded from the analyses; the attrition rate is 13 percent for the elementary school cohorts and 16 percent for the middle school cohorts.

Table 2. Correlations, Variances, Means, Sample Sizes, and Proficiency Rates for the Middle-School Cohorts

Grade	5	6	8	Variance	Mean	Low-Rigor Proficiency	High-Rigor Proficiency
Reference Cohort[a]					Sample Size = 252456		
5		.68	.66	8010	702	.85	.26
6	.73		.67	7868	735	.95	.50
8	.69	.75		8930	822	.94	.50
Variance	8118	8705	8104				
Mean	695	724	780				
Low-Rigor Proficiency	.88	.85	.84				
High-Rigor Proficiency	.41	.41	.25		Sample Size = 256598		
Prediction Cohort[a]					Sample Size = 271617		
5		.67	.63	7106	694	.86	.30
6	.74		.65	8118	739	.94	.46
8	.69	.76		8136	830	.92	.48
Variance	8299	8911	8082				
Mean	696	721	776				
Low-Rigor Proficiency	.87	.83	.83				
High-Rigor Proficiency	.42	.38	.24		Sample Size = 274241		

Data Analysis

The steps to examining predictive value are as follows. Each growth model is first used to classify each student as either on- or off-track. Next, students are isolated according to their performance levels (i.e., proficient and non-proficient) in grades 4 and 6, which are being treated nominally as the current grade levels for the elementary and middle school cohorts respectively. The research questions of interest are then answered by computing the positive predictive (PPV) and negative predictive (NPV) odds ratios as defined earlier in the paper. The number and percentage of students who fall within each condition of interest for the elementary and middle school cohorts are listed in Tables 3 and 4 respectively.

Table 3. Prevalence of Elementary School Cohort Students Maintaining, Declining, and Improving Performance

		Number (Percent) of Students				
Rigor	Subject	Overall	Maintaining Proficiency	Declining Performance	Maintaining Non-Proficiency	Improving Performance
Low	Reading	274600 (100)	229951 (84)	12715 (5)	17099 (6)	14835 (5)
	Mathematics	285649 (100)	239662 (84)	17600 (6)	18058 (6)	10329 (4)
High	Reading	274600 (100)	61896 (23)	18614 (7)	145547 (53)	48583 (18)
	Mathematics	285649 (100)	87612 (31)	22868 (8)	137773 (48)	37396 (13)

Table 4. Prevalence of Middle-School Cohort Students Maintaining, Declining, and Improving Performance

		Number (Percent) of Students				
Rigor	Subject	Overall	Maintaining Proficiency	Declining Performance	Maintaining Non-Proficiency	Improving Performance
Low	Reading	271617 (100)	241196 (89)	12973 (5)	9253 (3)	8195 (3)
	Mathematics	274241 (100)	207802 (76)	19742 (7)	27839 (10)	18858 (7)
High	Reading	271617 (100)	94109 (35)	29843 (11)	112174 (41)	35491 (13)
	Mathematics	274241 (100)	56272 (21)	48028 (18)	160620 (59)	9321 (3)

When generating a binary predictor such as an on-track indicator we are in essence transforming a predicted score into a dichotomy by choosing a threshold. The predictive accuracy of the binary predictor therefore depends on the choice of threshold. One way to examine the effect that the choice of threshold has on predictive accuracy would be to use several thresholds and report the results for each one. Computation of an ROC curve is an efficient method of doing just that (Ghonen, 2007).

The ROC curve has been widely applied in the medical and psychology fields (e.g., Begg, 1991; Green & Swets, 1966; Swets, 1973, 1988, 1996) to examine predictive quality. In constructing a ROC curve, predictions are expressed as binary on-track or off-track indicators across a range of thresholds (Green & Swets; Mason & Graham, 2002). For each threshold, the correspondence between the on-track/off-track indicators and observations of proficiency/non-proficiency in the target grade is examined. This correspondence is described by a two-component vector comprised of the 'hit rate' (the proportion of correctly predicted events) and the 'false-alarm rate' (the proportion of non-events that were incorrectly predicted as events). The hit rate and false-alarm rate are plotted in ROC space to describe the ROC curve (Mason & Graham).

For example, when analyzing PPV, the hit rate describes the sensitivity of the model, and the false alarm rate describes the type I error rate (i.e., 1 – specificity) at various thresholds of the predicted score. In addition to providing information about the error rates at various thresholds, the areas beneath the ROC curves (i.e., AUC) are also commonly used as summary measures of predictive quality (Ghonen, 2007; Mason & Graham, 2002; Sing, Sander, Beerenwinkel, & Lengauer, 2005).

Results

Results indicate that the predictive value of all growth models depends on the rigor of the target grade cut score placements. More precisely, the predictive value depends on the proficiency rate at the target grade. When predicting that students are on-track, the odds that a student will be proficient in the future given an on-track prediction get better as proficiency rates get higher.

The dependency of model results on proficiency rates was seen across both the elementary and middle school cohorts without evidence of interaction. Therefore, to help ease the results interpretation, only the middle school cohorts' results are displayed. The middle school cohorts were selected to display because they contain the extreme high and low target grade proficiency rates as well as a proficiency rate near 50 percent. Therefore, the distribution of target grade proficiency rates provides a clear picture of their effects on predictive value. The elementary school cohorts' results are similar enough to the middle school results that they add little of substance while increasing the amount of data to interpret. The complete set of results is available from the author upon request.

The odds that a student is proficient in grade 8 given an on-track prediction in grade 6 (i.e., PPV odds) are presented in Figure 2 for the student population as a whole. As depicted in Figure 2, we can be more certain that students predicted to be on-track will in fact be proficient as target grade proficiency rates increase, regardless of growth model used. This finding is not necessarily surprising. If almost all students are proficient in the target grade, then predicting that all current students are on-track will be very accurate. Nevertheless, there are differences in predictive value among the models.

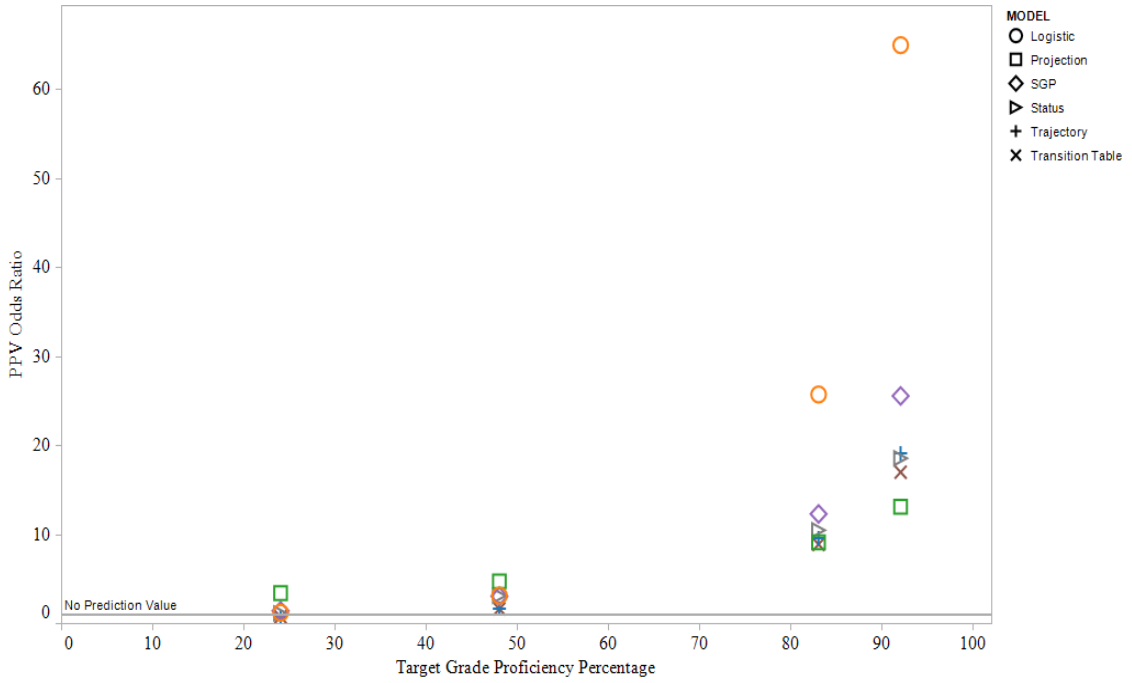

Figure 2. Odds of proficiency in grade 8 for all students given an on-track classification in grade 6 across growth models and target-grade proficiency rates. The "No Prediction Value" line reflects 1:1 odds.

Figure 3 displays growth model ROC curves for the grade 8 reading low-rigor condition, which has the highest target grade proficiency rate of all conditions, and the grade 8 mathematics high-rigor condition which has the lowest target grade proficiency rate. Note that all ROC curve figures presented in the paper display these same two conditions. Note also that the ROC curves were created for models that produce continuous predicted scores; therefore, the transition table model is not included on the ROC curves. However, the transition table produced very similar results to the trajectory model in this study, which supports previous research (Ho, Lewis, & Farris, 2009; Hoffer et al., 2011; Castellano & Ho, 2012). Therefore, the results seen for the trajectory model may be generalizeable to the transition table model.

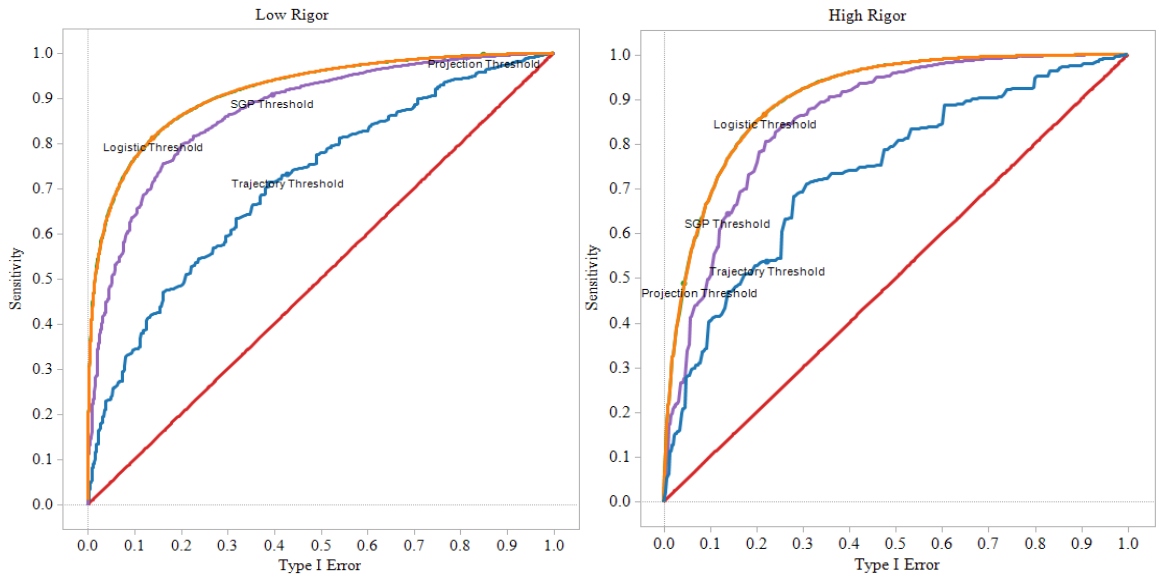

Figure 3. ROC curves and growth model on-track cut score thresholds for all students under low- and high-rigor conditions. The diagonal line represents no prediction value.

The first finding of note is that the models that assume average growth (i.e., projection and logistic regression) have larger AUC than the models that assume constant growth (i.e., trajectory and SGP). AUC is commonly used as a summary measure of predictive quality (Ghonen, 2007; Mason & Graham, 2002; Sing et al., 2005), suggesting that models assuming average growth have the potential for better predictive quality than models assuming constant growth. Of the models that assume constant growth, the SGP model displays larger AUC than the trajectory model. These findings are consistent throughout the more detailed analyses described later in the paper.

In addition, the logistic regression and projection models have almost identical ROC curves. Their performance is differentiated by where their on-track cut score thresholds are located on the curve. Under the low-rigor condition, the projection model threshold is located near the extreme levels of sensitivity and type I error rate, meaning it functions almost as a status model. By contrast, the logistic regression model's threshold

is located at 81 percent sensitivity and 14 percent type I error rate. This location describes the fact that the logistic regression model correctly classified 81 percent of the 249,391 proficient students as being on-track and only misclassified 14 percent of the 22,226 non-proficient students as being on-track under the low-rigor condition. Because the logistic regression model better discriminates students who maintain proficiency from those who do not, it produces better PPV.

Under the high-rigor condition, however, the projection model produces better PPV. The projection model's threshold is located at 49 percent sensitivity and 5 percent type I error rate, which means that the projection model correctly classified 49 percent of the 65,593 proficient students as being on-track and only misclassified 5 percent of the 208,648 non-proficient students as being on-track. Therefore, because more than three times as many students predicted to by on-track by the projection model are proficient as are non-proficient in the target grade, the PPV odds are higher than 3:1. On the other hand, the logistic regression model correctly classified 87 percent of the 65,593 proficient students as being on-track but misclassified 22 percent of the 208,648 non-proficient students as being on-track. As a result, approximately as many students predicted to be on-track by the logistic regression model are proficient as non-proficient in the target grade, and the odds of proficiency given an on-track classification are close to 1:1.

In summary, there is a tradeoff that must be negotiated between PPV, sensitivity, and type I error rate, and the proficiency rates in the target grade affect the manner in which that tradeoff must be negotiated as will be seen in the following sections.

Maintaining Proficiency

Figure 4 displays the PPV odds of maintaining proficiency given an on-track prediction across growth models and proficiency rates. As seen previously, the PPV odds depend on the target grade proficiency rate. Only the projection model offers predictive value when the proficiency rates are very low. The logistic regression model produces the best predictive value when proficiency rates are above 50 percent.

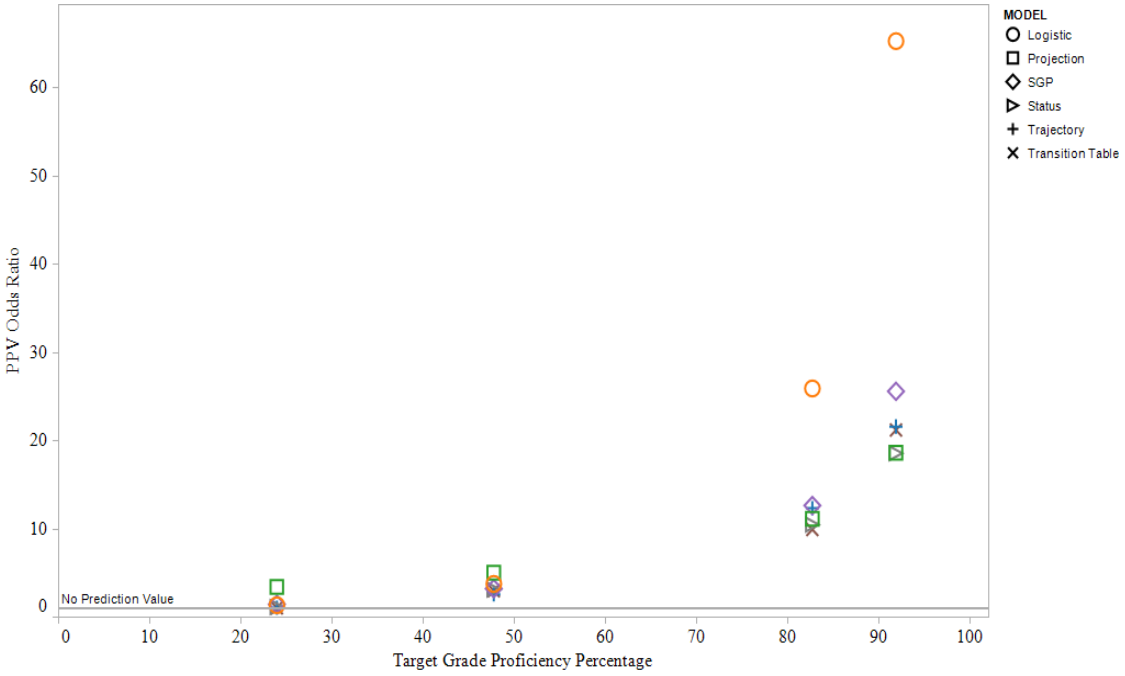

Figure 4. Odds that grade 6 students maintain proficiency in grade 8 given on-track predictions across growth models and target-grade proficiency rates. The "No Prediction Value" line reflects 1:1 odds.

Examination of the ROC curves in Figure 5 reveals that the logistic regression model provides the best balance among PPV, sensitivity and type I error rate under the low-rigor condition. By contrast, the projection model produces the best balance under the high-rigor condition. It is important to note that the AUC for all models' under the high-rigor condition is considerably smaller than under the low-rigor condition. This

suggests that the tradeoff among PPV, sensitivity, and type I error rate is more difficult to negotiate when predicting students will maintain proficiency and target grade proficiency rates are low.

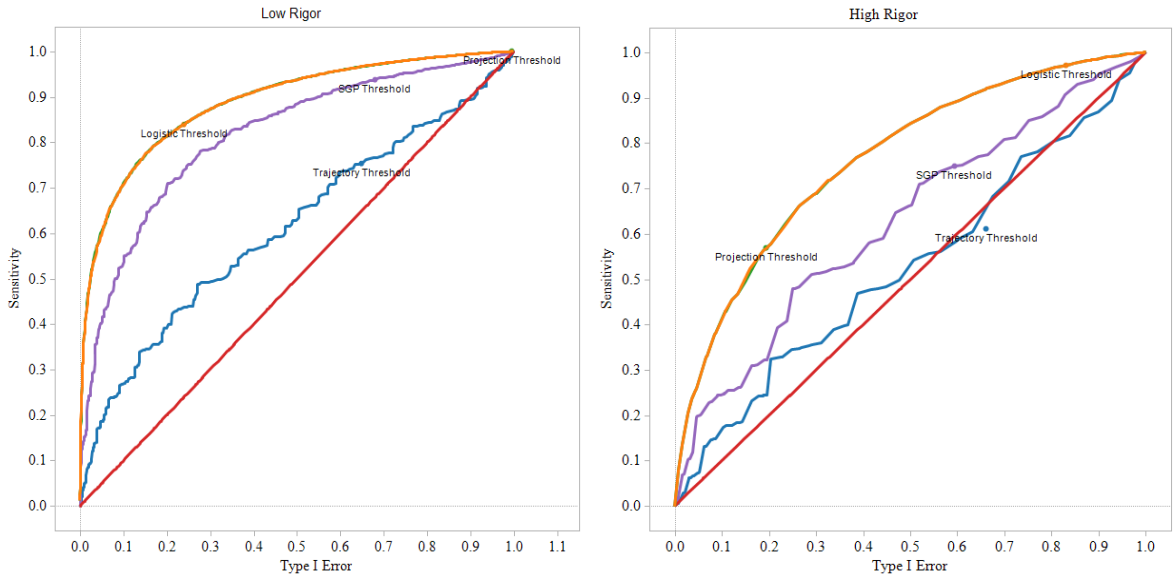

Figure 5. ROC curves and growth model on-track cut score thresholds for proficient grade 6 students under low- and high-rigor conditions. The diagonal line represents no prediction value.

Declining Performance

Figure 6 displays the odds of declining performance given an off-track prediction across growth models and proficiency rates. Only the logistic regression and projection models produce NPV odds that are higher than 1:1 across any of the target grade proficiency rates, and across three of the four proficiency rates those odds are only slightly higher than 1:1. The logistic regression model produces higher NPV odds when proficiency rates are below 50 percent, but the projection model produces higher odds when proficiency rates are above 50 percent.

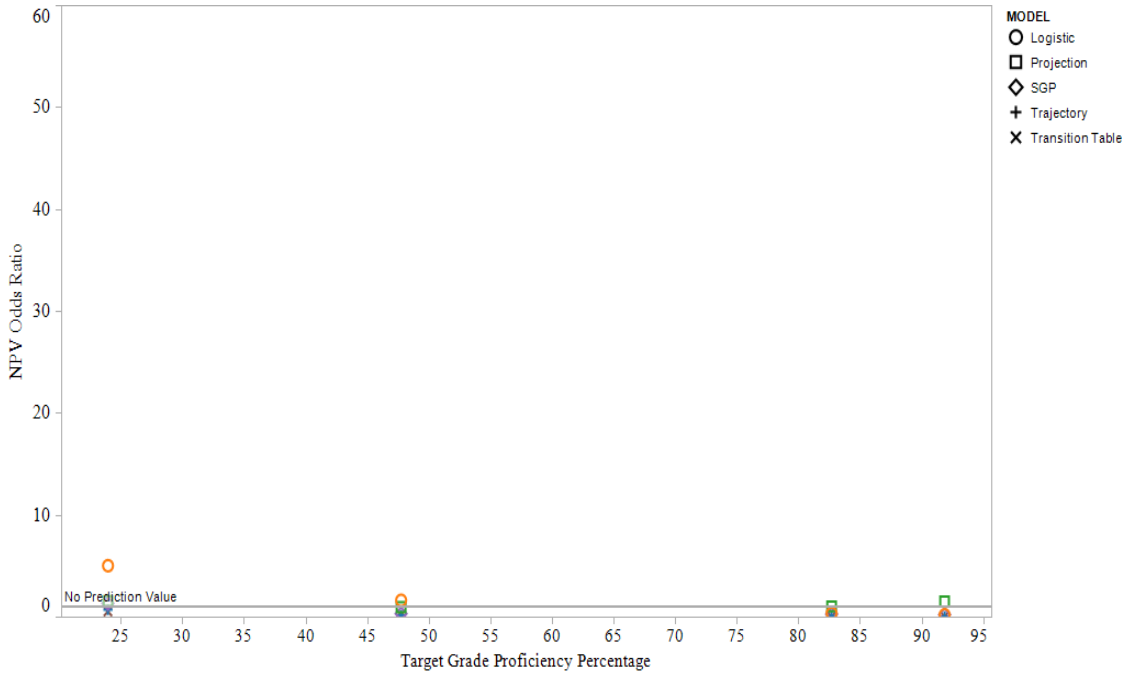

Figure 6. Odds that grade 6 students decline in proficiency in grade 8 given off-track predictions across growth models and target-grade proficiency rates. The "No Prediction Value" line reflects 1:1 odds.

Figure 7 displays the growth model ROC curves when predicting declining performance. Note that, because non-proficiency is the prediction of interest, specificity is on the *y* axis, and type II error rate (i.e., 1 – sensitivity) is on the *x* axis. The ROC curves provide interesting insights when viewed in combination with the NPV odds. First, the AUCs suggest that there may be some predictive power in regression based models under the low-rigor condition in particular. However, only the projection model produces odds that are higher than 1:1 under that particular condition (i.e., 8:5). Yet the threshold for the projection model is set at the bottom of the ROC curve near zero specificity. Specifically, the projection model correctly classified 0.45 percent of the 12,973 students who declined in performance as being off-track and misclassified 0.02 percent of the 241,196 students who maintained proficiency as being off-track. In other words, the on-track threshold for the projection model is set so low under this condition

that almost all students are projected to be on-track. The few students who get an off-track prediction are slightly more likely to be non-proficient than proficient in the target grade; however, the projection model misclassified a large proportion of the students who are truly declining in performance as being on-track.

By contrast, the logistic regression model correctly classified 76 percent of the 12,973 students who declined in performance as being off-track but misclassified 16 percent of the 241,196 students who maintained proficiency as being off-track. In other words, a cut score threshold that captures a significant proportion of the students who decline in performance under the low-rigor condition captures many more students who do not decline in performance.

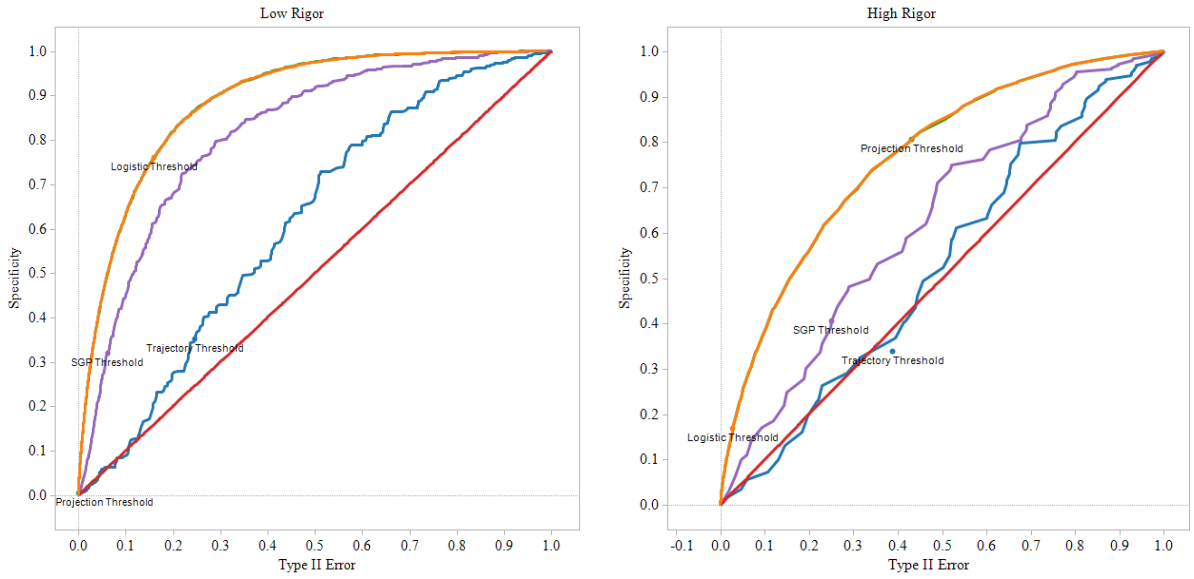

Figure 7. ROC curves and growth model off-track cut score thresholds for proficient grade 6 under low- and high- rigor conditions. The diagonal line represents no prediction value.

Under the high-rigor condition, the logistic regression and projection models offer better predictive value, but the tradeoff between specificity and type II error rate is still difficult to negotiate. Tracing the high-rigor ROC curve from the logistic regression

model threshold toward the projection model threshold in Figure 7 reflects decreasing NPV odds, increasing specificity, and increasing type II error rate. The logistic regression model correctly classified 17 percent of the 48,028 students who declined in performance as being off-track and misclassified 3 percent of the 56,272 students who maintained proficiency as being off-track. The NPV odds under the logistic regression model are therefore approximately 5:1. By contrast, the projection model correctly classified 81 percent of the 48,028 students who declined in performance as being off-track but misclassified 43 percent of the 56,272 students who maintained proficiency as being off-track. Therefore, the NPV odds under the projection model are only slightly higher than 1:1 (i.e., 8:5).

In summary, the projection and logistic regression models offer better predictive value than the SGP and trajectory models across both low- and high-rigor conditions. However, the predictive value that the models offer is limited, because very few of the students who are declining in performance are identified when NPV odds are substantially higher than 1:1, whereas identifying a substantial number of students who decline in performance results in NPV odds that offer little to no predictive value.

Maintaining Non-Proficiency

Figure 8 displays the odds of maintaining non-proficiency given an off-track prediction across growth models and proficiency rates. Examination of Figure 8 reveals that the NPV odds increase as the proficiency rates decrease. Thus, as target grade proficiency rates decrease, we can be more certain that students who are classified as off-track will in fact be non-proficient at the target grade regardless of growth model. The logistic regression model displays the best NPV odds when proficiency rates are low, but the projection model displays the best NPV odds when proficiency rates are high.

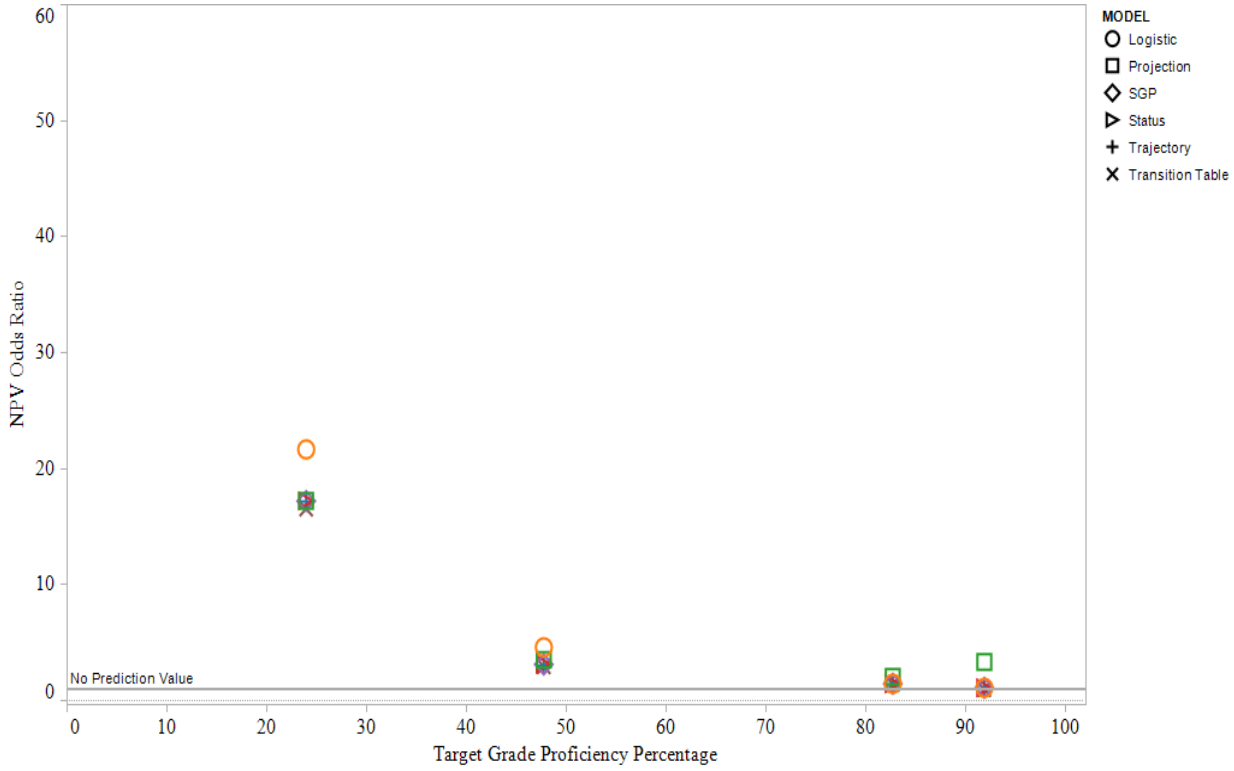

Figure 8. Odds that grade 6 students maintain non-proficiency in grade 8 given off-track predictions across growth models and target-grade proficiency rates. The "No Prediction Value" line reflects 1:1 odds.

Figure 9 displays the growth models' ROC curves. Under the low-rigor condition, the logistic regression, SGP, and trajectory models function similarly to a status model in

that they all classify almost all non-proficient grade 6 students as being off-track. On the other hand, the projection model produces better predictive value (i.e., NPV odds higher than 3:1) but lower specificity (i.e., 36 percent).

Under the high-rigor condition, the ROC curves indicate that the growth models offer little beyond a status model that assumes currently non-proficient students will remain non-proficient in the future. Although the logistic regression model offers more predictive value than the others, all models produce high type II error rates, meaning they misclassified a large proportion of students who improved from non-proficiency to proficiency as being off-track.

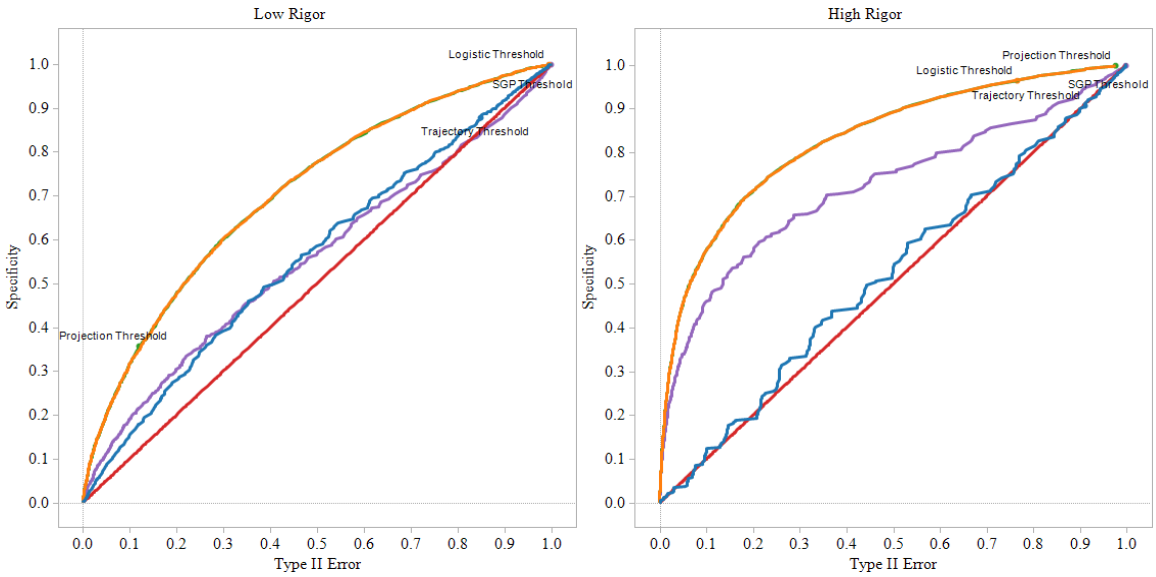

Figure 9. ROC curves and growth model off-track cut score thresholds for non-proficient grade 6 students under low- and high-rigor conditions. The diagonal line represents no prediction value.

Improving Performance

Examination of Figure 10 reveals that the PPV odds of improving performance depend on the target grade proficiency rate. None of the models offers predictive value

when the proficiency rates are very low. The logistic regression model produces the most predictive value when proficiency rates are above 50 percent.

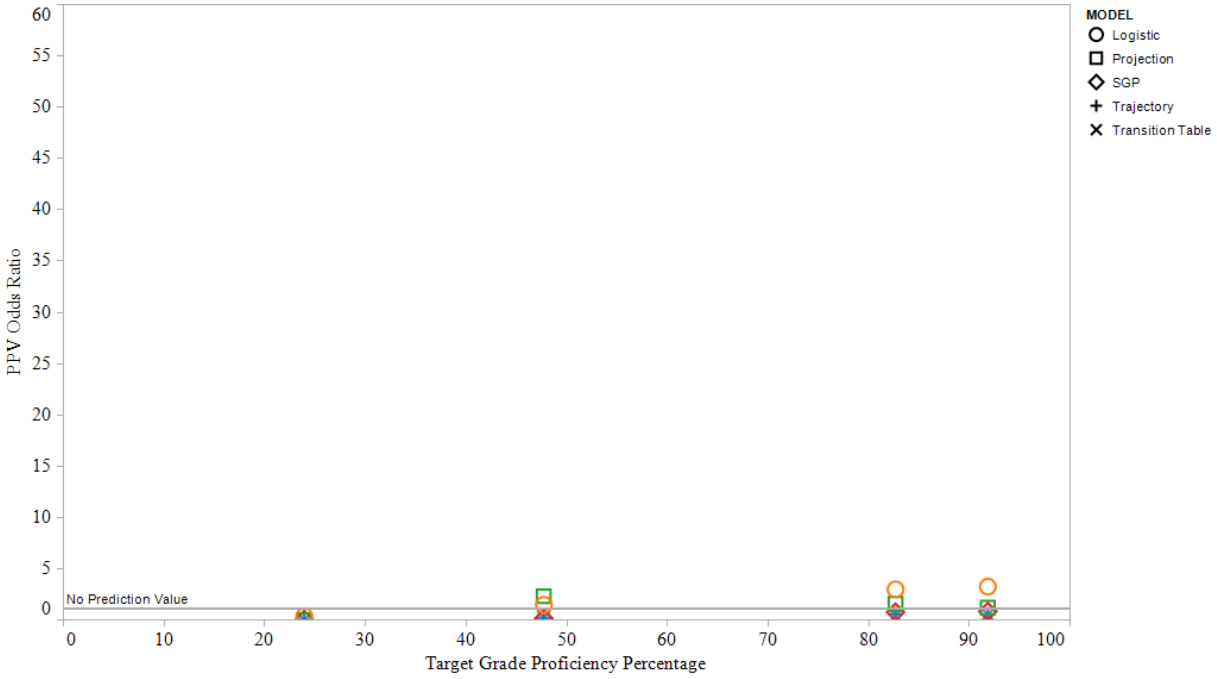

Figure 10. Odds that grade 6 students improve in proficiency in grade 8 given on-track predictions across growth models and target-grade proficiency rates. The "No Prediction Value" line reflects 1:1 odds.

Examination of the ROC curves in Figure 11 reveals that there is a tradeoff among PPV odds, sensitivity, and type I error rate that has no elegant solution across any of the examined conditions when predicting improving performance. To produce high PPV odds as the logistic regression model does under the low-rigor condition, a cut score threshold must be set at a score such that students who attain it are more likely than not to be proficient in the future. However, the majority of students who actually improved their performance did not meet such a rigorous threshold and were misclassified as being off-track, which reduces predictive value. On the other hand, reducing the rigor of the threshold, as the projection model does under the low-rigor condition, results in the

misclassification of substantial proportions of students who maintained non-proficiency as being on-track, which also reduces predictive value. It is important to note that this tradeoff is similarly difficult to negotiate throughout all of the examined conditions, including those where relatively large numbers of students improved their performance (e.g., see high-rigor reading Tables 3 and 4).

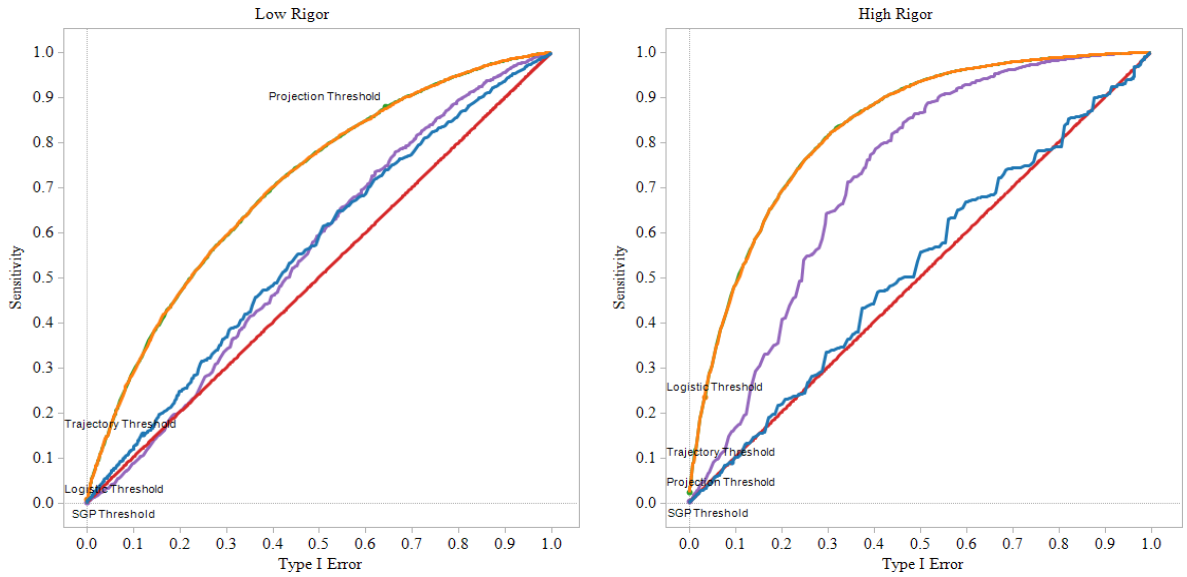

Figure 11. ROC curves and growth model on-track cut score thresholds for non-proficient grade 6 students under low- and high-rigor conditions.

Discussion

There are several important findings resulting from this study. First, it appears that models relying on an assumption of average cohort growth provide more predictive value than models relying on an assumption of constant individual growth. Therefore, their use is recommended to set growth targets or assign early-warning indicators. This finding may also have important implications for the SGP model. The results of this study suggest that the SGP model may add predictive value when assessing whether

students will "catch up" or "keep up" by assuming that all students will progress toward a future outcome at the 50th percentile rate of growth, rather than assuming that individual students will progress at a rate corresponding to their most recent student growth percentile calculation.

Second, the results of this study suggest that the models are not good at assigning on-track indicators to identify student progress that suggests movement from the current performance level to a different performance level in the future. None of the models was able to discriminate well between students who are improving or declining in performance and students who are maintaining their performance level. This may have less to do with the models themselves than with the use of binary on-track/off-track indicators.

The poor growth model predictive value is not because students do not improve or decline in their performance across time – they do. The lack of predictive value stems from the fact that these trends are somewhat anomalous, meaning students who improve or decline in performance have similar achievement profiles to students who do not. As a result, type I and type II errors are inevitable, and the prevalence of each type of error will depend on the target grade proficiency rate and the location of the cut score threshold. Type I errors could result in students not receiving necessary interventions; type II errors could result in students receiving unnecessary interventions.

Therefore, it may be prudent to avoid binary classifications such as on-track/off track when predicting growth. Predicted scores may be more useful if they enable teachers to make decisions according to how far the predicted scores are from the target grade cut scores. Teachers may want to design instruction differently, for example,

depending on whether students are predicted to be far below, near, or far above the target grade cut score. One potential drawback to this approach is teachers may misinterpret the precision of the predicted scores, because they may not be familiar with the statistical concept of prediction error. Logistic regression is an attractive alternative if this is a concern. The results of this study indicate that the logistic regression model offers the same predictive quality as a projection model, and it has some potential advantages. Logistic regression models report scores in terms of probabilities, which may be a more intuitive method than confidence intervals for communicating to stakeholders the inherent uncertainty that accompanies growth predictions.

Alternatively, a third category could be added as a means of highlighting that not all on-track predictions are equally likely. For example, students with predicted scores that are well above the target grade cut score could be categorized as safely on-track (e.g., "green light"); students with predicted scores that are in a region where type I and type II errors are likely could be categorized with a cautionary early-warning indicator (e.g., "yellow light"); students with predicted scores that are well below the target grade proficiency score could be categorized with an early-warning indicator noting that significant intervention is warranted (e.g., "red light"). Logistic regression and ROC analyses can provide information about type I and type II errors at various prediction cut score thresholds which can inform the selection of cut scores that separate the three categories.

To summarize, growth prediction models that predict future proficiency based on average cohort growth may provide actionable information to educators, but reporting outcomes in terms of binary on-track indicators should be avoided if possible. Logistic

regression models report outcomes in terms of probabilities, which are easy to understand and therefore an attractive alternative. Reporting outcomes in terms of predicted scale scores or categorical indicators that consist of more than two categories may also be useful if teachers are mindful of prediction error.

Finally, it is important to note that the results of this study are based on the data from one state's standardized assessment. Therefore, the extent to which the results generalize to statewide assessments with different psychometric properties is unknown. In addition, the models tested in this paper are generic and not necessarily reflective of models that are used in practice. It is possible that using more sophisticated methods to specify the growth prediction models would lead to different results.

References

Begg, C. B. (1991) Advances in statistical methodology for diagnostic medicine in the 1980s. *Statistics in Medicine*, *10***,** 1887–1895.

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

Bewick V., Cheek L., & Ball J. (2004). Statistics review 13: Receiver operating characteristic (ROC) curves. *Critical Care, 8***,** 508–512.

Briggs, D. C. (2011). Making Inferences about Growth and Value-Added: Design Issues for the PARCC Consortium. *A White paper Commissioned by the Partnership for Assessment of Readiness for College and Careers consortium.* Retrieved from http://www.parcconline.org/sites/parcc/files/BriggsPARCCGrowthFINAL022412.pdf.

Castellano, K. E., & Ho, A. D. (2012). A Practitioner's Guide to Growth Models. Retrieved from http://scholar.harvard.edu/files/andrewho/files/castellano_and_ho_-_practitioners_guide_to_growth.pdf

Colorado Department of Education. 2008. *The Colorado growth model: Higher expectations for all students.* Retrieved from www2.ed.gov/admins/lead/account/growthmodel/co/cogrowthproposal101508.pdf.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal*, *47*(4), 458-472.

Gonen, M. (2007); *Analyzing Receiver Operating Characteristic Curves Using SAS*, SAS Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Peninsula Publishing, Los Altos, California, USA.

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.

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145-2166.

Partnership for Assessment of Readiness for College and Careers. (2010). *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.

Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal detectability: Part I. The general theory*. Electronic Defense Group, Technical Report 13, June 1953. Available from EECS Systems Of. ce, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940-3941.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240***,** 1285–1293.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*, Lawrence Erlbaum Associates, Mahwah, NJ.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182***,** 990–1000.

U.S. Department of Education (2006). Letter to Chief State School Officers reminding of the opportunity to participate in the Department's Growth Model Pilot for the 2006–07 school year. Retrieved from http://www.ed.gov/policy/elsec/guid/secletter/061011.html.

U.S. Department of Education (2005). *Secretary Spellings announces growth model pilot* [Press Release]. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html.

Version, S. A. S. 9.2 SAS Institute Inc. *SAS Campus Drives, Cary, NC, 27513*.

Washington State, & SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program application for new grants*.

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

Weiss, M. J. (2008). *Using a yardstick to measure a meter: Growth, projection, and value-added models in the context of school accountability*. (Doctoral dissertation), Available from UMI. (3309569) Retrieved from http://proquest.umi.com/pqdlink?did=1537009541&Fmt=6&VType=PQD&VInst=PROD&RQT=309&VName=PQD&TS=1343067805&clientId=79356.

Weiss, M. J., & May, H. (2012). A policy analysis of the federal growth model pilot program's measures of school performance: The Florida case. *Education Finance and Policy, 7(1),* 44-73.