The Dependence of Growth-Model Results on Proficiency Cut Scores

The Dependence of Growth-Model Results on Proficiency Cut Scores

Andrew D. Ho

University of Iowa

Daniel M. Lewis and Jason L. MacGregor Farris

CTB/McGraw-Hill

Abstract

States participating in the Growth Model Pilot Program reference individual student growth against "proficiency" cut scores that conform with the original No Child Left Behind Act (NCLB). Although achievement results from conventional NCLB models are also cut-score dependent, the functional relationships between cut-score location and growth results are more complex and are not currently well described. We apply cut-score scenarios to longitudinal data to demonstrate the dependence of state- and school-level growth results on cut-score choice. This dependence is examined along three dimensions: 1) rigor, as states set cut scores largely at their discretion, 2) across-grade articulation, as the rigor of proficiency standards may vary across grades, and 3) the time horizon chosen for growth to proficiency. Results show that the selection of plausible alternative cut scores within a growth model can change the percentage of students "on track to proficiency" by more than 20 percentage points and reverse accountability decisions for more than 40% of schools. We contribute a framework for predicting these dependencies, and we argue that the cut-score dependence of large-scale growth statistics must be made transparent, particularly for comparisons of growth results across states.

Keywords: growth models, standard setting, NCLB

The Dependence of Growth-Model Results on Proficiency Cut Scores

U.S. Secretary of Education Margaret Spellings introduced the Growth Model Pilot Program (GMPP) in November of 2005 (U.S. Department of Education, 2005). The GMPP encourages the incorporation of individual student growth into the accountability calculations of the No Child Left Behind Act (NCLB). Growth models are based on longitudinal growth in individual achievement as opposed to the cross-sectional student status that anchors conventional NCLB calculations. Whereas conventional NCLB models recognize students crossing a particular proficiency cut score, growth models offer states the potential to reward growth across a broader range of achievement.

The GMPP was introduced to provide greater flexibility to states challenged with meeting the goals of NCLB. The U.S. Department of Education indicated that "[t]he purpose of this pilot is 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, 2006a). GMPP regulations thus constrain the architecture of growth models to accountability calculations that support current NCLB principles. Specifically, GMPP requirements indicate that "[t]he accountability model must ensure that all students are proficient by 2013-2014" (U.S. Department of Education, 2006b). Progress toward universal proficiency, a primary goal of NCLB, must thereby remain the reference point for accountability calculations (U.S. Department of Education, 2006c).

The NCLB focus on universal proficiency has been criticized for allowing unintended consequences and inaccurate interpretations of achievement (Rothstein, Jacobsen, & Wilder, 2006). For example, the variability of proficiency standards across states has confounded high student achievement with lenient proficiency standards (Braun & Qian, 2007; Linn, 2003;

McLaughlin & Bandeira de Mello, 2005). Holland (2002) and Ho (2007) have shown that proficiency-based trends and gap trends are subject to surprising changes in magnitude and sign under alternative proficiency cut scores. And NCLB's proficiency-based accountability framework may encourage the disproportionate allocation of school resources to students who are just below the proficiency cut score (Booher-Jennings, 2005; Neal & Schanzenbach, 2007). Although the GMPP allows for alternate approaches to determine AYP, it may also add an additional layer of complexity to proficiency-based calculations and increase the likelihood of misinterpretation of student, school, and state progress toward proficiency.

In this paper, we describe how GMPP-based accountability results are dependent on a) attributes of the cut scores adopted by states and b) attributes of growth model policies themselves. Attributes of cut scores include features such as rigor—where more rigor implies higher cut scores across all grades and less rigor implies lower cut scores across all grades—and articulation, where the rigor of cut scores may vary systematically across grades. We also review attributes of growth model policies, and we focus specifically on the impact of the chosen time horizon to proficiency. For example, schools may receive credit for nonproficient students who are on track to proficiency in 3 years but not those who are on track to proficiency in 4 or more years. These two sets of attributes are responses to two different waves of federal policy. Cut-score decisions were made largely in response to the basic requirements of NCLB, and growth model policy decisions were made largely in response to the GMPP. Before describing the impact of these attributes on growth model results, we discuss each of these sets of attributes and their policy contexts in turn.

Attributes of Cut Scores: Rigor and Across-Grade Articulation

When growth is referenced to cut scores, growth model results will depend upon both the rigor of cut scores and the patterns of cut score articulation across multiple grades. The decision-making process that leads to these attributes has many motivations. Prior to NCLB implementation, most states tested at several benchmark grades, and only a few states tested contiguously across all or even most elementary and middle school grades (Olson, 2002). The contiguous-grade testing paradigm that emerged under NCLB necessitated setting cut scores for the newly required assessments and, in many cases, resetting cut scores previously set without regard to NCLB consequences. Two issues were addressed by state policymakers as they considered designs for setting or resetting cut scores on their NCLB assessments: the rigor of the cut scores within each grade and the articulation of cut scores across the grades.

Efforts to foster consistency of cut scores across grades were only moderately successful under the benchmark-grade testing paradigm, where standard setting tended to be a grade-by-grade activity. Substantial differences between cut scores for Grades 4 and 8, for example, are easier to reconcile than inconsistencies between cut scores for adjacent grades. The contiguous-grade testing paradigm impelled policymakers and standard-setting researchers to develop methods to vertically moderate, or articulate cut scores across grades (Cizek, 2005). These methods were guided by the principle that the pattern of proficiency rates across grades should appear rational to the various constituents who use or interpret test results. The pattern of cut scores in each grade should consider the scope and sequence of content across the grades and should reflect the regular progress that students tend to achieve from grade to grade.

There are no prescribed standards for what constitutes consistent across-grade results in terms of the percentage of students at or above a given performance standard. However, several patterns of well articulated performance standards are frequently observed in state assessment

programs as a result of their explicit consideration in the standard setting process. These models describe articulated, cross-grade performance standards in terms of the percentage of proficient students in each grade. Three general models and interpretations of articulated across-grade performance standards are presented by Lewis and Haug (2005).

The decreasing model reflects a smoothly decreasing percentage of proficient students across grades. There are multiple interpretations of this pattern of across-grade proficiency. It may reflect the changing nature of the domain across grades and, with it, a decrease in students' ability to meet the goals of the grade. For example, mathematics becomes increasingly complex, moving from the reasonably simple and concrete notions of counting, addition, and subtraction to the complex and abstract foundations of algebra and calculus. Lewis and Haug (2005) also observe that standard-setting participants in upper grades are often content-area experts who may act as gatekeepers of the domain, whereas elementary teachers tend to be more student centered, considering what is best for the student and hesitating to label younger students as less than proficient.

The equal-percentage model reflects an equal or approximately equal percentage of proficient students in each grade. This pattern reflects an attribute of proficiency that is inherent in many performance-level descriptors—that proficient students are well prepared to meet the challenges of the next grade. Thus, proficient students tend to meet the challenges necessary to demonstrate proficiency in their next grade, resulting in similar percentages of proficient students from one grade to the next.

The increasing model reflects a smoothly increasing percentage of proficient students across grades. There are also multiple interpretations of this pattern of across-grade proficiency. It may be that a somewhat higher bar is set at the lower grades to increase the likelihood that

students will be prepared for the more challenging material—and possibly higher stakes assessments—that come in subsequent grades. Additionally, as students' strengths and weaknesses are better understood through appropriate longitudinal record keeping, teachers and parents may provide better educational opportunities customized to individual learning styles, leading to an accelerated success rate with time.

This paper frames both the rigor and the across-grade consistency of cut scores as factors influencing the results of growth models under the GMPP. Rigor is investigated by increasing cut scores in all grades from those that would set proficiency rates near 90% in all grades (low rigor) to those that would set proficiency rates near 30% in all grades (high rigor). Articulation is investigated by setting cut scores that enforce increasing percentages of proficient students (decreasing rigor) across grades by 10 percentage points per grade, then 9, then 8, and so on. This rate of decline can be set to 0, an equal-percentage model, and then a rate of decreasing across-grade proficiency rates (increasing rigor) is investigated. By varying cut-score attributes in a plausible range and holding other factors constant, we demonstrate that rigor and articulation can change growth percentages by up to 20 percentage points, and school-level accountability results can be even more dramatically affected. In the next section, we discuss attributes of growth model policies that have similarly large influences on growth results under the GMPP.

Attributes of Growth Model Policies

Under conventional NCLB accountability models, a student is classified as "proficient" at Time t if the student's test score, X_t, is greater than or equal to a cut score designating proficiency at that time, c_t. Table 1 displays the conditions for student proficiency as well as other classifications soon to be introduced. The cornerstone statistic of NCLB is the percentage of proficient students, denoted here as PPS. The PPS is calculated for each sufficiently large

subgroup in a school and compared to a benchmark called the Annual Measurable Objective (AMO). While so-called safe-harbor provisions and, for some states, idiosyncratic confidence-interval procedures may complicate matters, the baseline NCLB rule is that a school's PPS statistics must be greater than the AMO for all valid subgroups in order to avoid sanctions.

Schools whose subgroups all have PPS statistics higher than the AMO are described as making Adequate Yearly Progress (AYP, Table 1). In a simplified scenario with only one subgroup, this decision process may be represented by the statement: If PPS $\geq$ AMO, then AYP. Using this terminology, proficiency describes a student, PPS is a school-level percentage, and AYP describes a school. At the state-level, relevant statistics include the PPS, which can be calculated for a state as well as for a school, and the percentage of schools making AYP (PAYP). The PPS at both the state- and school-level is inversely related to the rigor of a cut score: the higher the cut score, the lower the PPS. The PAYP is more complex, as it is dependent on the AMO. However, for fixed AMOs, PAYP will also decline with increasing cut-score rigor. This should seem fairly straightforward: School PPS will decline as cut-score rigor increases, and fewer school PPS will surpass the AMO. We will demonstrate that the cut-score dependencies of growth statistics are just as predictable but much less straightforward.

To date, there are 11 growth models with full or conditional approval by the Department of Education (U.S. Department of Education, 2008). Many of these growth models afford the classification of students as "on track" to proficiency at some point in the future. We denote the percentage of "on track" students in a school or a state as POT. A state's growth model policy may count "on track" students as "proficient" for the sake of school accountability decisions; we adopt this policy throughout our analyses. We identify schools whose accountability classification is reversed by using growth models: We count the schools that are not making

AYP, that is, PPS < AMO, but have surpassed the AMO through their "on track" students, that is, PPS + POT ≥ AMO. We distinguish these schools from conventional AYP schools and describe them as having made Adequate Yearly Growth (AYG, Table 1).

A state's growth model policies can dictate the definition of POT and the form of the inequality that decides AYG. For some states, current status is subtracted from the equation entirely, and students must be predicted to be on track to proficiency whether they are currently proficient or not. Under this system, there is no PPS; there is only POT, and a school makes AYP if POT ≥ AMO. States may also use models that predict future scores from prior test scores using regression-type methods. These models use longitudinal data from students from prior years to estimate prediction equations. Other states may award fractional credit to on-track students based on the starting point and/or degree of their gains. In order to focus on the functional relationships between select variables, we use a simplistic model that avoids confounding the multiple policy factors that make up AYP and AYG decisions in practice. Dunn (2008, this issue) describes the effects of the policies of approved growth model states as they work *in vivo*, whereas we demonstrate how all states can expect results to change under systematic manipulation of select factors.

The model we use for illustration, described in Table 1, is commonly referred to as a gain-score or trajectory growth model. Student scores must be located on a vertical scale that spans many grades, and cut scores are also mapped onto this scale. The model assumes that a student's gain over a past unit of time will be the same as that student's gain over each similar unit of time into the future. For example, a nonproficient student whose trajectory is projected from Time 1 and 2 scores of 425 and 475, respectively, is on track to 525 at Time 3, 575 at Time 4, and so on. The growth model policy, then, defines this nonproficient student as "on track" to

proficiency at Time 3 if the projected score of 525 is above the Time 3 proficiency cut score. This can be represented by the following conditions: $X_2 < c_2$ and $X_1 + 2*(X_2 - X_1) \geq c_3$ (Table 1). In a straightforward extension of this logic, at Time 2, students are defined as "on track in N years" if the following inequalities hold: If $X_2 < c_2$ and $X_1 + (N + 1)*(X_2 - X_1) \geq c_{N+2}$.

This paper investigates how a particular growth model policy attribute—the time horizon to proficiency, N—affects accountability decisions. We demonstrate that cut-score attributes can interact with growth model policy attributes to dramatically affect growth statistics for states (POT) and for schools (PAYG). We show that increasing the time horizon to proficiency can more than double the percentage of students who are classified as "on track." In addition, we explain how changing cut scores can interact with AMOs with the potential to reduce school failure rates by more than 35 percentage points. These policy decisions are among many that states must make in implementing a growth model, but we demonstrate that these particular decisions have systematic relationships with outcome variables and are easy to predict. We conclude this paper with an argument for the generalizability of these dependencies to all growth models that reference growth to NCLB-type proficiency cut scores.

Methods: A Theoretical Framework for Evaluating Cut-Score Dependencies

Visualization and prediction of the cut-score dependence of growth results can be assisted by a theoretical framework. The cornerstone of the theoretical framework is a bivariate scatterplot with student scores at Time 2 (X_2) plotted against student scores at Time 1 (X_1). We present an illustration of observed test score data generated from a bivariate normal distribution in Figure 1. The bivariate normal distribution was chosen for convenience and because many tests have distributions that are either scaled to be or happen to be unimodal and roughly symmetrical. Departures from bivariate normality will certainly change the findings presented.

However, our purpose here is illustrative, and the framework easily allows for alternative distributional choices or, as we will show, examples with real data.

Figure 1 shows a sample of 1000 students drawn from this bivariate normal distribution and plotted as small gray dots. The scale is arbitrary up to a linear transformation; the mean of the Time 1 distribution is set to 500, and the standard deviations of the Time 1 and Time 2 distributions are set to 100. The Time 2 mean is set to 550, resulting in an average gain of 0.5 standard deviation units, a level commonly seen in practice. The correlation between Time 1 and Time 2 scores is also set to be realistic at 0.75.

To establish a reference point, this framework assumes that the current year is Time 2, and the first year of data, Time 1, came from the previous year. The centroid is indicated by a black circle. The $X_2 = X_1$ diagonal is plotted for reference. Points above this diagonal represent students whose scores have increased from Time 1 to Time 2, and points below the diagonal represent students whose scores have decreased from Time 1 to Time 2. For illustration, Figure 1 flags two students' data; one student is represented as a triangle and one as a square. The triangle identifies a student who scored 350 at Time 1 and 475 at Time 2 for a gain of 125. The gain can be visualized as the vertical or, equivalently, horizontal distance from any point to the diagonal. The square identifies a student who has a Time 1 score of 425 and a Time 2 score of 475 for a gain of 50. This student has a smaller gain and is closer to the diagonal. We will demonstrate how this framework readily identifies which students are classified as "on track."

Five reference lines have been drawn across each axis. Each line marks the cut scores set at Times 1, 2, 3, 4, and 5; these are labeled c_1, c_2, c_3, c_4, and c_5, respectively. In Figure 1, the cut scores are set at 450 at Time 1, 500 at Time 2, 550 at Time 3, and so on. Of the two flagged students, the triangle is below the Time 1 cut score on the horizontal axis and also below the

Time 2 cut score on the vertical axis. In other words, this student is below proficient in both years. The student has also made a gain of 125 points from Time 1 to Time 2. If the student were to make the same gain from Time 2 to Time 3, the student would score a 600, which is above the Time 3 cut score. The student represented by the black triangle is therefore on track to proficiency in 1 year.

The student represented by the black square has a Time 1 score of 425 and a Time 2 score of 475 for a gain of 50. Note that this student is also below proficient in both years. If the student makes the same gain from Time 2 to Time 3, the student will score a 525, which is not proficient in Time 3. This student is not on track to proficiency in 1 year. The usefulness of this framework is that students who are "on track" to proficiency can be easily identified by the area of the graph in which they are located. The lightest shaded area identifies the students who are "on track in 1 year" and is bordered by the horizontal line: $X_2 = c_2$, and the diagonal line: $X_2 = (X_1 + c_3)/2$. The area below the horizontal line identifies nonproficient students at Time 2. The area above the diagonal line identifies students who have made gains that place them on track to proficiency by Time 3. The equation for the diagonal line follows directly from the equation in Table 1 after solving for X_2. The percentage of students in this area are what the gain-score model in Table 1 would describe as POT_1. Note that the triangle has no border on its left side and is therefore semi-infinite.

Figure 1 also shows the successively darker shaded triangles that identify the students on track in 2 years (but not 1) and 3 years (but not 2 or 1), respectively. The full percentage of students on track in 3 years, POT_3, is the sum of the proportions calculated from all three shaded triangles. The equation of each diagonal line can be found by solving the general equation in Table 1 for X_2 given a particular time horizon of N years. The figure shows that increasing the

time horizon to proficiency allows for greater and greater proportions of "on track" students under this growth model.

If the bivariate normal model adequately describes the distribution of observed scores, POT statistics can be calculated as the volume under the density function in the region of the semi-infinite triangles shown in Figure 1. The appropriate double integral for the first triangle, representing the percentage of students on track for proficiency in 1 year, follows:

$$POT_1 = \int_{-\infty}^{2c_2-c_3} \int_{\frac{X_1+c_3}{2}}^{c_2} f(X_1, X_2)\, dX_1 dX_2 \qquad (1)$$

Here, $f(X_1, X_2)$ is the usual bivariate normal density function with five parameters:

$$f(X_1, X_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 + \left(\frac{X_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2 - 2\rho\left(\frac{X_1-\mu_{X_1}}{\sigma_{X_1}}\right)\left(\frac{X_2-\mu_{X_2}}{\sigma_{X_2}}\right)\right]\right] \qquad (2)$$

In the scenario shown in Figure 1, the means would be 500 and 550, the standard deviations would both be 100, the correlation would be 0.75, and c_2 and c_3 would be 500 and 550 respectively. We calculate the integral using numerical quadrature in the program, Matlab, and we find that POT_1 under these parameters would be 4.71%. From here, it is straightforward to adjust cut scores in Equation 1 and recalculate POT under alternative cut score selections. For example, the cut scores shown in Figure 1 could all be shifted much lower or much higher to represent less or more rigor respectively. A generalized version of Equation 1 allows evaluation of percentages of students on track to proficiency in N years:

$$POT_1 = \int_{-\infty}^{\frac{(N+1)c_2 - c_{N+2}}{N}} \int_{\frac{Nx + c_{N+2}}{N+1}}^{c_2} f(X_1, X_2)\, dX_1 dX_2 \qquad (3)$$

The score scale shown in Figure 1 is hypothetical, and the implication of a Time 2 cut score of 500, for example, is difficult to discern on its own. A more interpretable approach to describing cut scores references cut scores by the proficiency statistics they generate. For example, a Time 1 cut score of 450 and a Time 2 cut score of 500, as shown in Figure 1, result in PPS of around 69% for both grades. To explain the effects of the rigor of cut scores on growth statistics, we begin by assuming that cut scores follow an equal-percentage model across grades, that is, the percentage of proficient students is the same in all grades. Distributions in higher grades are assumed to have equal standard deviations of 100 and equal average gains of 50 per year for convenience in calculating cut scores. We then shift the entire set of cut scores so that the percentages of proficient students in all grades ranges from 90% (low cut scores and less rigor) to 30% (high cut scores and more rigor), a plausible range that spans the PPS statistics seen in practice (Swanson, 2008). Visually, this is akin to taking the grid in Figure 1 and sliding it up and down the diagonal of the graph, calculating the proportion of students in the areas of the triangles while the grid moves.

Figure 2 shows results from calculations that vary both overall rigor of cut scores from 90% (less rigor) to 30% (more rigor) and the time horizon to proficiency from $N = 1$ to 5 years. To maintain consistency with the framework presented in Figure 1, we order the horizontal axis by the increased rigor that comes from raising cut scores, thus the proficiency rates of figures will decrease from left to right. Figure 2 shows that low levels of rigor lead to low POT statistics

of around 2%, whereas increasing rigor can increase POT statistics to 10% when the time horizon is 1 year and to over 20% when the time horizon is 5 years.

Most time horizons for GMPP states are 3 years, though time horizons may be shorter for students near graduation from a particular school or when the year approaches 2014, the deadline for 100% proficiency. States that have equal time horizons but dramatically different levels of rigor may report POT differences of more than 15 percentage points due simply to cut score selection. It may seem counterintuitive that increasing rigor leads to greater percentages of students who are on track. The simple explanation for this is that increasing rigor results in more nonproficient students and thus more students eligible to qualify for growth calculations. Visually, this can be pictured in Figure 1, where increasing cut scores will result in greater and greater proportions of students bordered by the semi-infinite shaded triangles.

Figure 2 has striking implications for comparisons of growth-model results across states. It is a reminder that the percentage of students who are "on track" (POT) is confounded with cut-score selection. States reporting high PPS are expected to have relatively lower POT statistics, and states with more rigorous cut scores are expected to experience the greatest benefit from the GMPP under projection growth models. These benefits would increase if the time horizon to proficiency were to increase, though the differences between 3-year and 4-year time horizons are not as dramatic as those between 1-year and 2-year time horizons.

Real Data Confirmation of Theoretical Relationships

The relevance of the theoretical framework was evaluated by applying similar cut-score scenarios to longitudinal student data from a mid-sized state. The dataset consists of a single-grade cohort of almost 70000 students who have four years of test scores from Grades 3 through 6 inclusive. There are over 83000 eligible scores in Grade 6, so the match rate for the four-year

span is approximately 84%. The test is an English Language Arts test that is vertically scaled with increasing grade means and decreasing variability over the four grades.

To calculate the cut-score dependence of POT_1, the percentage of students on track to proficiency in 1 year, Grade 4 and Grade 5 scores are defined as X_1 and X_2 respectively. The cut scores c_1, c_2, c_3, are set by a similar equal-percentage model as the one that generated Figure 2. For example, for a target PPS of 80%, empirical cut scores are identified that result in 80% of students classified as proficient at each grade. The cut scores associated with a given target percentile are denoted c_1, c_2, and c_3, and the target PPS is then varied to investigate cut-score dependencies. To calculate POT_2, Grades 3 and 4 are defined as X_1 and X_2 respectively, and c_1, c_2, c_3 and c_4, are calculated as the percentiles of the empirical distributions of Grades 3, 4, 5, and 6. With a four-year dataset, only two different time horizons can be evaluated: a 1-year horizon using Grade 4-5 growth projected to Grade 6 and a 2-year horizon using Grade 3-4 growth projected to Grade 6. Similar results could be obtained for a 1-year horizon using Grade 3-4 growth projected to Grade 5 but were not included for the sake of parsimony.

Figure 3 shows the results of varying cut scores to reflect equal cross-grade PPS from 90% to 30% in a similar manner to Figure 2. As in Figure 2, the POT_1 line runs from near 1% for less rigorous cut scores to near 10% for more rigorous cut scores. The POT_2 is slightly higher than the corresponding line in Figure 2 and shows greater cut-score dependence. Looking at Figure 1, this can be explained by a greater density of students in the second semi-infinite triangle. As the real data has a slightly higher correlation and a slightly lower gain than the theoretical data, there is indeed a greater density of students in the region of the second triangle, leading to the observed results. The curves in Figure 3 are bumpy because of the usual awkwardness arising from calculating percentiles from discrete data. The similarities between

Figures 2 and 3 support an argument for the consistent functional dependencies between cut-score choice and POT statistics.

Effects of Articulation of Cut Scores: Theoretical and Real Data Results

Cut scores may decline or increase in rigor across grades. In order to model the effects of the articulation of cut scores across grades, we introduce a series of plausible PPS patterns in Table 2. These patterns all have an average PPS of 60% across all grades, but higher patterns in the table show declining rigor across grades (increasing PPS) and lower patterns show increasing rigor across grades (decreasing PPS). The central pattern is an equal-percentage pattern that generates 60% proficiency at each grade. The cut scores that would establish these PPS statistics are calculated for the theoretical score distributions underlying Figure 1. For example, for the first pattern in Table 2, the cut scores that generate the listed percentages are approximately 513, 537, 561, and 583. The integrals in Equations 1 or 2 can then be evaluated for cut scores generated from each of the patterns in Table 2. In Figure 1, this exercise can be visualized by constricting the distance between the cut scores (for decreasing rigor) and then expanding the distance between the cut scores (for increasing rigor). As the distances between cut scores increases, the diagonal lines in Figure 1 are pulled up and squeezed against horizontal line $X_2 = c_2$. As a result, the proportion of students in the area of the triangle is decreased if rigor increases across grades. Moving from the top rows of Table 2 to the bottom rows can be seen as an exercise in increasing the distances between cut scores.

The results of the evaluation of these integrals for time horizons of 1 and 2 years are shown in Figure 4. Simply put, raising future goals will decrease the number of students who are on track to these goals. Figure 5 confirms these findings with the real dataset previously described. Similar to the contrast between Figures 2 and 3, Figures 4 and 5 are very similar for a

1-year horizon, but the empirical data for the 2-year horizon shows a greater dependence on cut-score attributes than the theoretical data. This may also be explained by a disproportionate weight of students in the "on-track-in-2-years" triangle.

The magnitudes of the dependencies in Figures 4 and 5 are slightly less dramatic than those shown in Figures 2 and 3. Further, the extremes shown in Figures 4 and 5 are slightly less realistic, as declines or increases in proficiency rates of 30 percentage points across four grades are not common in practice. In contrast, PPS ranges between 90% and 30%, as shown in Figures 2 and 3, represent the actual PPS variation currently seen across states (Swanson, 2008). This seems to suggest that the practical range of cut-score rigor has a greater impact than the practical range of cut-score articulation. It is nonetheless impossible to fully disentangle the these two factors in longitudinal analysis, as increasing distances between cut scores naturally affects rigor in each grade. These interactions can be visualized in Figure 1 through the stretching and shifting of the grid of cut scores over the density of data in the scatterplot. Together, Figures 2-5 demonstrate that increasing the overall rigor of cut scores increases POT statistics, and increasing rigor from grade to grade decreases POT statistics.

School-Level Accountability Results

To this point we have described the effects of cut-score rigor, cut-score articulation, and time horizons on state-level results for students, as represented by the percentage of on-track students (POT) statistic. A more relevant statistic for some policymakers may be the percentage of schools for whom growth models may make a difference in accountability decisions. A school-level version of the theoretical framework in Figure 1 exists, however the number of variables and interdependencies becomes too complicated for the framework to support helpful

visualizations. Instead, we simply show the empirical results for the dependence of the proportion of AYG schools on the rigor of cut scores.

The matched dataset contains information for approximately 70000 students in over 900 elementary schools. The data are longitudinal and stretch over four years for a single-grade cohort from Grade 3 to Grade 6. For the purposes of stability, and in order to mimic the minimum subgroup size for this state and many others, we exclude all schools whose available matched data number fewer than 30 students. This leaves around 640 schools for a 71% school-inclusion rate.

AYP decisions are made on multiple subgroups within schools, where all PPS must surpass AMOs. For simplicity, we consider the single-grade cohort as the only subgroup in the school. Additionally, we only use the 2-year time horizon, where the "current" Grade 4 includes growth results from Grades 3 to 4 and credits students on track to proficiency by Grades 5 or 6. Finally, we did not include safe-harbor or confidence-interval provisions whose interactions with growth models further complicates dependencies. The implications of these findings to schools with multiple subgroups can proceed by referencing the PPS of the lowest-scoring subgroup, as schools are essentially accountable to this subgroup alone. The school-level accountability decision, which is effectively a model for Grade 4 in our scenario, follows from Table 1: If PPS $\geq$ AMO, then AYP, and, for growth model decisions for non-AYP schools: If PPS + POT $\geq$ AMO, then AYG.

AMOs are essentially cut-scores used to determine whether individual schools have met their AYP goals. As such, they have an effect on growth model results that is comparable to the impact of the student-level proficiency cut-score choice. As NCLB took effect, the AMO was tied to PPS such that states were discouraged from setting excessively low standards for schools.

Our purpose is to demonstrate the dependence of PAYG on proficiency cut scores while fixing other factors, but it is unrealistic to model the shifting of cut scores without a corresponding shift in the AMO. To appropriately model AMO correspondence to a given cut-score , we follow the federal formula that originally linked a state AMO to a state PPS. NCLB requirements set the minimum AMO at the PPS of "the school at the 20th percentile in the State, based on enrollment, among all schools ranked by the percentage of students at the proficient level" (Pub. L. No. 107-110, 2002). This amounts to the following algorithm: For each grade and its associated proficiency cut score, 1) Rank all schools by their PPS; 2) Calculate the cumulative enrollment as a percentage of statewide enrollment from the lowest ranked school on up; and 3) Set the AMO equal to the PPS of the school at which 20 percent of the statewide enrollment is reached. As before, we generate sets of cut scores that lead to equal PPS across grades. Following the algorithm above, each cut score sets a PPS which in turn determines an AMO.

Figure 6 illustrates the dependence of school-level growth model results on the rigor of cut scores by extending the student-level results of Figure 3 to schools. Only the 2-year time horizon is shown, and AMOs are recalculated for each PPS using the calculation described above. The vertical axis represents PAYG, the percentage of schools that do not meet AYP but achieve AYG through the growth model. The darker, black line shows the dependence of PAYG on cut-score rigor. We find that increasing rigor increases the proportion of AYG schools. This is similar to the student-level results shown in Figure 3, where increasing rigor also increases the proportion of "on track" students. Reducing the proficiency rate from 90% to 30% increases PAYG from 10% to over 20%.

The lighter, gray line reflects the result of adding 5 percentage points to the AMO, effectively raising the minimum standard for schools while keeping all other data constant.

Under NCLB and the GMPP, all AMOs are required to rise from their baseline values to 100% by 2014. The gray line illustrates the results of an increase in the AMO with no corresponding change in state proficiency rates. While this would decrease PAYP, Figure 6 shows that this would increase PAYG by 5 to 10 percentage points. The gray line's position over the black line shows that, if AMOs increase over time and proficiency rates stagnate, the impact of growth models on school accountability decisions may become even greater.

It may seem surprising that PAYG never dips below 10%. Figure 6 seems to suggest that all growth model states should have shown differences for at least 10% of schools after implementation. Instead, many states observed changes at only a handful of schools (Klein, 2007). A more complete picture of the interaction between AMOs and cut-score rigor may help to resolve the apparent conflict between Figure 6 and real-world findings. While the AMO was federally mandated at the advent of NCLB, it has since become uncoupled with PPS. PPS statistics track observed student achievement annually while AMO trajectories are set by state policy, increasing from their baseline values to 100% in 2014. When AMOs and PPS become uncoupled, even more dramatic dependencies can manifest.

Figure 7 displays the results for fixed AMOs of 50% and 70%, levels that represent points along most state AMO trajectories towards 100% by 2014. The results show that the impact of a growth model on a state's schools can be both very large and deeply dependent on cut-score attributes. A solid line at the 65% mark illustrates a scenario where a state sets cut scores such that 65% of its students are proficient across all grades. For this state, the GMPP would positively affect 11% of schools if the state AMO were 50% and 31% of schools if the state AMO were 70%. Figure 7 shows that states will have particularly large growth-model benefits when the PPS is just below the AMO. In these cases, a large number of schools will be

on the AMO bubble, and adding POT to the calculation results in a larger proportion of AYG schools. If proficiency rates rise faster than the baseline AMO, Figure 7 shows that PAYG is expected to be quite low. For example, if 75% of students are proficient, and the AMO is set to 50%, the empirical results from this state show only 5% of schools making AYG.

Together, Figures 6 and 7 allow the following observations about the potential impact of the GMPP on states as a function of state PPS levels and its AMO. First, states whose PPS far exceeds their AMO are likely to experience little benefit from the GMPP. Second, the impact of the GMPP will be greatest, as measured by the peaks in Figure 7, when both the PPS and AMO are a) similar and b) in the middle range of percentages. When both the PPS and AMO are large, PPS suppresses POT (see Figures 2 and 3) and thus suppresses PAYG. Third, for states whose AMO trajectories rise to meet and then surpass their PPS levels, the impact of the GMPP will rise and then fall. This latter finding will be realized by many states should NCLB reach its endgame. All of these observations are best described as straightforward consequences of a cut-score-based growth model and not as meaningful differences in amounts of growth across states or over time.

Generalizing Findings to Alternative Growth Model Policies

The number of variables involved in an operational state growth model is far too large to explore all possible interactions between cut-score attributes and policy decisions. To this point, we have described how cut-score attributes affect one particular growth model approach—the gain-score or trajectory growth model—in combination with a time-horizon factor. In this section, we briefly discuss how the cut-score dependencies of this model may or may not generalize to three alternative implementations of growth models.

The growth model described to this point can only help students and schools. The inequalities displayed in Table 1 leave all proficient students in place and can only add "on track" students to school accountability calculations. An alternative formulation may choose to penalize proficient students who are not on track to proficiency. Under this formulation, all students, regardless of their status, must be making gains that show them as "on track." The student-level and school-level equations reduce to: If $X_1 + 2*(X_2 - X_1) \geq c_3$, then "on track," and, if POT $\geq$ AMO, then AYG. In this model, proficiency and AYP are irrelevant as long as students have growth data.

This model can be visualized in Figure 1 by extending the diagonal lines through to the other side of the $X_2 = X_1$ diagonal. Everyone above these lines is "on track in N years," and everyone below these lines is not. The net change between a growth model and a conventional status model must incorporate both the addition of the shaded triangles already highlighted in Figure 1 and the subtraction of new triangles bordered by $X_2 = c_2$ and the diagonal lines on the right side of the graph. These new triangles include currently proficient students who are not on track to proficiency and who would be classified as effectively nonproficient under the terms of this growth model.

It is clear that the growth model would no longer have a purely positive effect, but we argue that cut-score dependencies would certainly remain. Keeping triangles on both sides of Figure 1 in mind, we can see that low cut scores would actually lead to a net negative impact on states, as many proficient students would be classified as "not on track." Higher cut scores would allow the positive effects of growth models to become more salient. Thus, growth models that apply to all students regardless of their current proficiency status will be expected to have a less positive impact but remain dependent on cut-scores.

The second alternative policy model that we consider is regression-based. Regression-based models use data from previous longitudinal cohorts to generate prediction equations. If a student's scores from a current cohort are substituted in to the prediction equation and the equation returns a score above a future cut score, that student may be deemed "on track."

Regression-based models are still interpretable within the framework of Figure 1. The diagonal lines in Figure 1, for example, the "on track in 1 year" line: $X_1 + 2*(X_2 - X_1) = c_3$, are of exactly the same form as a regression-based prediction line. In fact, if all of the parameters of the multivariate normal model stayed the same, and the Time 1 to Time 3 correlation were set to 0.3 (admittedly a low value), the regression-based model would return a line with identical slope and an intercept around 15 points below the "on track in 1 year" line in Figure 1. Resulting cut-score dependencies would take on a similar form as the ones we have shown here. The impact of a regression-based growth model on POT depends critically upon the parameters of the distributions and thus the slope of this prediction line. Under different parameters, the regression line will change in slope, but its intercept will remain referenced by the future proficiency cut score. The framework in Figure 1 will still apply: raising cut scores will still leave more students to be classified as "on track," and raising future cut scores (increasing cut-score variability) will continue to decrease POT by raising the intercept of the regression line. Cut-score dependencies may therefore take on a different form but are not rendered negligible by regression-based growth models.

The third alternative policy we consider are value tables or categorical growth models. These use multiple cut-scores within a given grade to classify students, for example, into Below Basic, Basic, Proficient, and Advanced categories. Student transitions across category boundaries may receive some form of credit for schools. This model may also be visualized in

Figure 1. Instead of diagonal lines, categorical growth models add greater numbers of vertical and horizontal lines corresponding to the cut scores separating, for example, the Below Basic and Basic category in each grade. Instead of shaded triangles above the main diagonal that identify students receiving credit, categorical growth models will have shaded rectangles that are weighted by certain values.

Categorical growth models will therefore exhibit similar patterns of cut-score dependence as their gain-score model counterparts. Time horizons generally do not apply in categorical growth models. Cut-score articulation becomes a more complex concept as multiple cut-scores interact within and across grades, however raising higher-grade standards with respect to lower-grade standards will still decrease POT. Finally, increasing cut scores increases the number of nonproficient students who will be included in the rectangles. Across all of these alternative policy models, Figure 1 helps to demonstrate that cut scores will have a strong and often confounding effect on growth-based classifications and decisions.

Discussion and Conclusions

This paper has provided a framework for the quantification of the cut-score dependence of growth model results. We point out that the proportion of students credited by growth models should be larger for states with more rigorous cut scores primarily because of the increased proportion of nonproficient students eligible for growth calculations. We also show that this credited proportion is smaller for states whose distances between cut scores increase up the vertical scale (declining PPS across grades) and larger for states whose distances between cut scores decrease up the vertical scale (increasing PPS across grades). We show that the expected increase in credited students under different time horizons is much larger moving from a 1-year to a 2-year horizon than it would be moving from a 4-year to a 5-year horizon. Finally, we

demonstrate that the proportion of schools that will be benefitted by a growth model will generally increase with more rigorous cut scores. However, the effects ultimately depend on an interaction between cut scores and the state AMO: States with PPS close to their AMO will have the greatest proportion of benefited schools, particularly when that AMO is not too high.

Under the GMPP, more rigorous cut scores result in greater proportions of students and schools credited as "on track" and meeting federal goals. In this way, the GMPP provides a modicum of balance to the challenges states face under NCLB. Policymakers adopted cut scores with a tension between national reform efforts demanding high standards and NCLB's challenging goal to have universal proficiency by 2014; more rigor satisfies the former while less rigor supports achievement of the latter. The result has been a wide range of proficiency results across states that is partially if not mostly explained by differences in rigor (Braun & Qian, 2005; McLaughlin & Bandeira de Mello, 2003). States adopting more rigorous cut scores have a greater baseline challenge than states adopting less rigorous cut scores. Our results indicate that the use of this type of growth model offsets the difference between the challenges facing these states.

However, the findings of this study indicate that the interpretation of GMPP statistics as reflecting "growth," per se, is inaccurate or at least incomplete. Two hypothetical states with the same baseline level of student achievement, that adopt the same growth model at the same time, and that experience the same increase in student achievement over time, will experience different proportions of students and schools classified as "on track" as a direct result of levels of rigor, different patterns of vertical articulation of performance standards, or different time horizons. Consequently, growth results are only as comparable across states as their cut-score rigor, articulation, and time horizons are comparable across states.

The interpretation of NCLB results would be enhanced had cut scores been more consistent across states in the baseline year—relative progress toward meeting standards would acquire a common frame of reference. This did not occur, and across-state comparisons of results that rely on this assumption may encourage substantive conclusions about state differences. Instead, these differences are more appropriately attributed to the cut-score features that generated them. Interpreting the differences illustrated in Figures 2 through 7 as meaningful differences in student growth for different states would be a mistake, given that the underlying bivariate score distributions were exactly the same. It is the dependence of the growth-to-proficiency metric on cut-score attributes that produced the result. This degree of conflation between cut scores and growth results may be an undesirable attribute for models adopted under federal educational accountability systems.

Policymakers and standard setters should be aware that decisions they make about proficiency standards have a direct, dramatic, and, as we show, predictable impact on growth results in a growth-to-proficiency framework. For example, in 2001, Colorado adopted cut scores with an approximately equal percentage of proficient students across grades for its NCLB Reading assessment, whereas, in 2004, Indiana adopted Reading cut scores such that rigor increased and the percentage of proficient students decreased across grades. The results of this paper indicate that the adoption of a projection growth model would likely result in more schools classified as making AYG in Colorado than in Indiana, although the relative rigor of the cut scores would moderate the effects.

To conclude, our concerns are twofold. First, the cut-score dependencies demonstrated here are substantial. Statistics of interest like POT and PAYG can swing by dramatic amounts over plausible alternative cut scores and time horizons. These findings should be seen as parallel

to but contrasting with those of Allen, Briggs, Weeks, and Wiley (2008, this issue). They address the impact of scaling and linking methods, whereas we address cut-score and policy attributes. The factors that they investigate can be seen in Figure 1 as influencing the bivariate scatterplot, whereas the factors we investigate influence the grid overlaying the bivariate scatterplot. Clearly, both sets of factors will have an impact on policy-relevant outcomes.

Second, reporting growth from within a proficiency-based and therefore cut-score-dependent framework makes it difficult to satisfy two important requirements: transparency and parsimony. The degree of cut-score dependence is too substantial to disentangle growth from cut-score attributes and still allow for growth-related interpretations. This does not mean that growth-to-proficiency models should be thrown out as a possible tool of policy. However, it does require an honest recasting of results not as measures of growth but as indicators of progress toward a very particular standard.

One strategy for defensible interpretations involves the separation of growth-based accountability and growth-based reporting onto two parallel tracks. Toward the latter goal, a more straightforward method of encouraging accurate growth interpretations may follow from setting norm-referenced standards for growth (Betebenner, 2008; this issue). As these efforts progress, this paper stands as a reminder that the impact of the GMPP on state accountability metrics will be heavily moderated by cut-score attributes in systematic and predictable ways.

References

Allen, J., Briggs, D., Weeks, J., & Wiley, E. (under review for this special issue). The impact of vertical scaling decisions on growth projections.

Betebenner, D. (under review for this special issue). Normative and criterion referenced conceptions of student growth.

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

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313-338). New York, NY: Springer.

Dunn, J. (under review for this special issue). Holding schools accountable for the growth of non-proficient students: Coordinating measurement and accountability.

Cizek, G. J. (Ed.). (2005). Special issue of Applied Measurement in Education on the topic of Vertically Moderated Standards, *18*(1): Lawrence Erlbaum Associates, Inc.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, *26*(4), 11-20.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*, 3–17.

Klein, A. (2007). Impact is slight for early states using 'growth'. *Education Week*. *27*(16), 24-25.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education*, *18*, 11-34.

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

McLaughlin, D., & Bandeira de Mello, V. (2005). *How to compare NAEP and state assessment results*. Presented at the 35th Annual National Conference on Large-Scale Assessment. Retrieved April 18, 2007, from http://38.112.57.50/Reports/LSAC_20050618.ppt

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

No Child Left Behind Act of 2001. Pub. L. No. 107-110, 20 U.S.C § 6301 et seq. (2002).

Olson, L. (2002, January 9). Testing systems in most states not ESEA-ready. *Education Week*, *21*(16), pp. 1, 26-27.

Rothstein, R., Jacobsen, R., & Wilder, T. (2006, November). 'Proficiency for all' – An oxymoron. In, *Examining America's commitment to closing achievement gaps: NCLB and its alternatives*. Symposium conducted at the meeting of the Campaign for Educational Equity, New York, NY. Retrieved April 18, 2007, from http://www.epinet.org/webfeatures/viewpoints/rothstein_20061114.pdf

Swanson, C. B. (2008, January 10). Grading the states. *Education Week*, *27*(18), 36-38.

U.S. Department of Education. (2005, November 18). Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond. *U.S. Department of Education Press Release*. Retrieved February 12, 2007, from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

U.S. Department of Education. (2006a, October 11). 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 (October 11, 2006) Retrieved August 28, 2008, from http://www.ed.gov/policy/elsec/guid/secletter/061011.html

U.S. Department of Education. (2006b, January 25). *Peer Review Guidance for the NCLB Growth Model Pilot Applications*. Retrieved January 15, 2008, from http://www.ed.gov/policy/elsec/guid/growthmodelguidance.pdf

U.S. Department of Education. (2006c, May 17). *Summary by the Peer Review Team of April 2006 Review of Growth Model Proposals.* Retrieved January 15, 2008, from http://www.ed.gov/admins/lead/account/growthmodel/cc.doc

U.S. Department of Education. (2008, June 10). *U.S. Secretary of Education Margaret Spellings Approves Additional Growth Model Pilots for 2007-2008 School Year*. Retrieved August 28, 2008, from http://www.ed.gov/news/pressreleases/2008/06/06102008.html

Table 1. A simplified accountability model for evaluating cut-score dependencies of student- and school-level accountability classifications at Time 2. Assumes a single subgroup per school, two years of student scores (Time 1 and 2), and cut scores on a vertical scale.

Classification	Definition	Abbreviations for Percentages
Student-Level Classifications		
Proficient	If $X_2 \geq c_2$.	Percentage of Proficient Students (PPS)
"On Track" to proficiency in 1 year	If $X_2 < c_2$ and $X_1 + 2*(X_2 - X_1) \geq c_3$.	Percentage of On Track Students (POT_1)
"On Track" to proficiency in *N* years	If $X_2 < c_2$ and $X_1 + (N + 1)*(X_2 - X_1) \geq c_{N+2}$.	Percentage of On Track Students (POT_N)
School-Level Classifications		
Adequate Yearly Progress (AYP)	If PPS ≥ AMO.	Percentage of AYP Schools (PAYP)
Adequate Yearly Growth (AYG)	If PPS < AMO and PPS + POT ≥ AMO.	Percentage of AYG Schools (PAYG)

Legend: X_t is a student score at Time *t*.

c_t is the proficiency cut score at Time *t*.

Figure 1. Theoretical growth framework with shaded areas indicating students "on track" to proficiency in 1, 2, and 3 years.

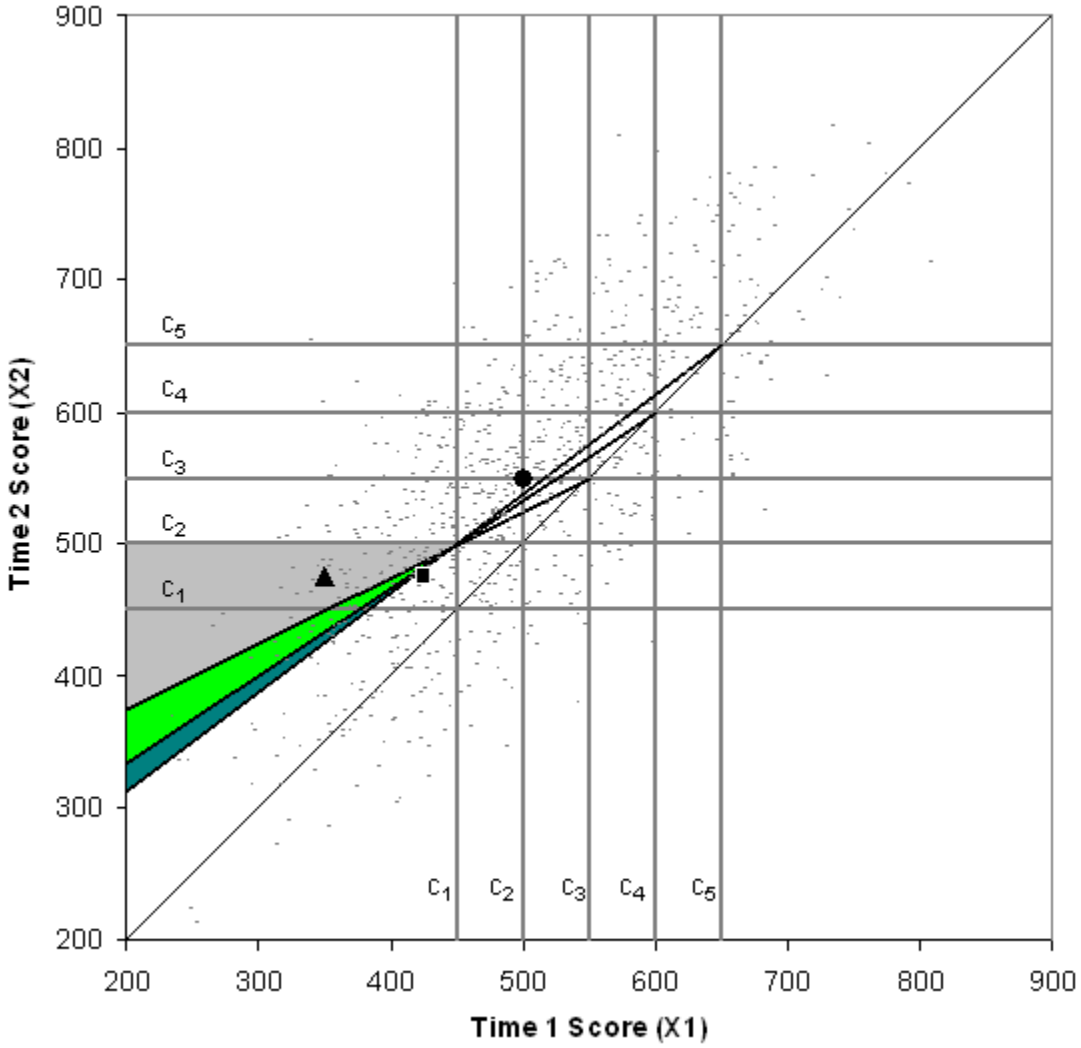

Figure 2. Theoretical dependence of the percentage of "on track" students on cut-score choice and time horizon to proficiency.

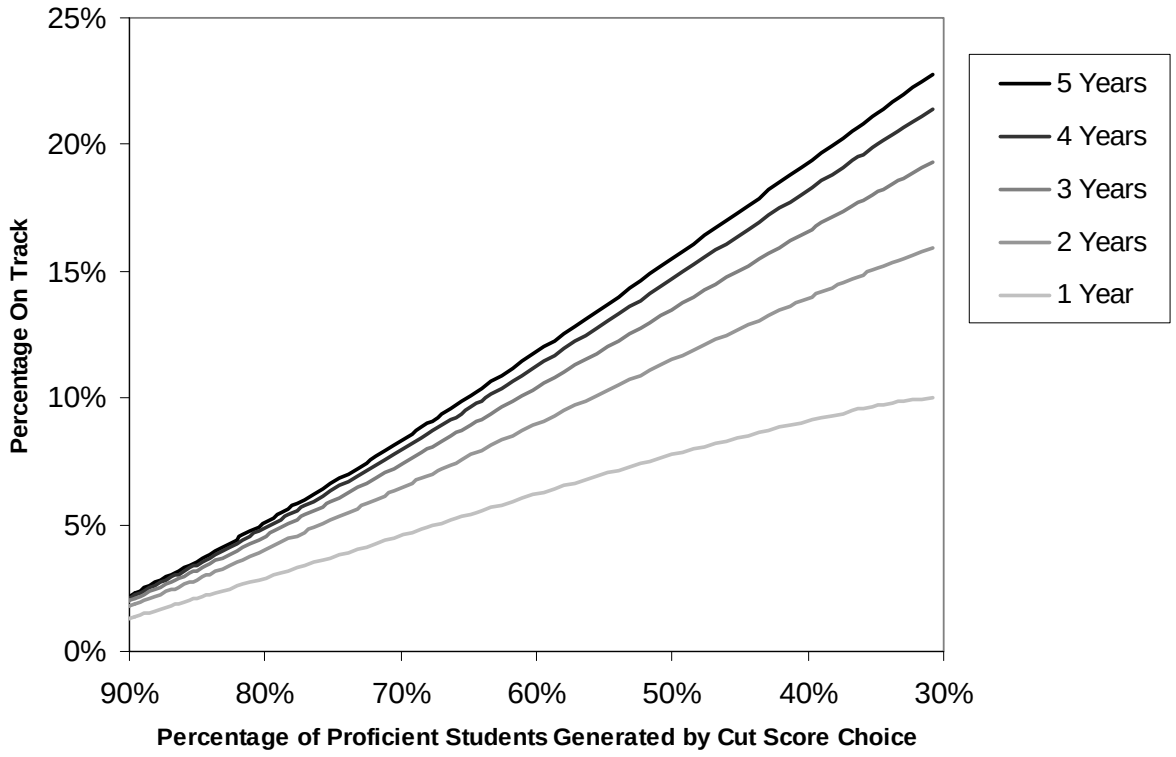

Figure 3. Empirical dependence of the percentage of "on-track" students on cut-score choice and time horizon to proficiency.

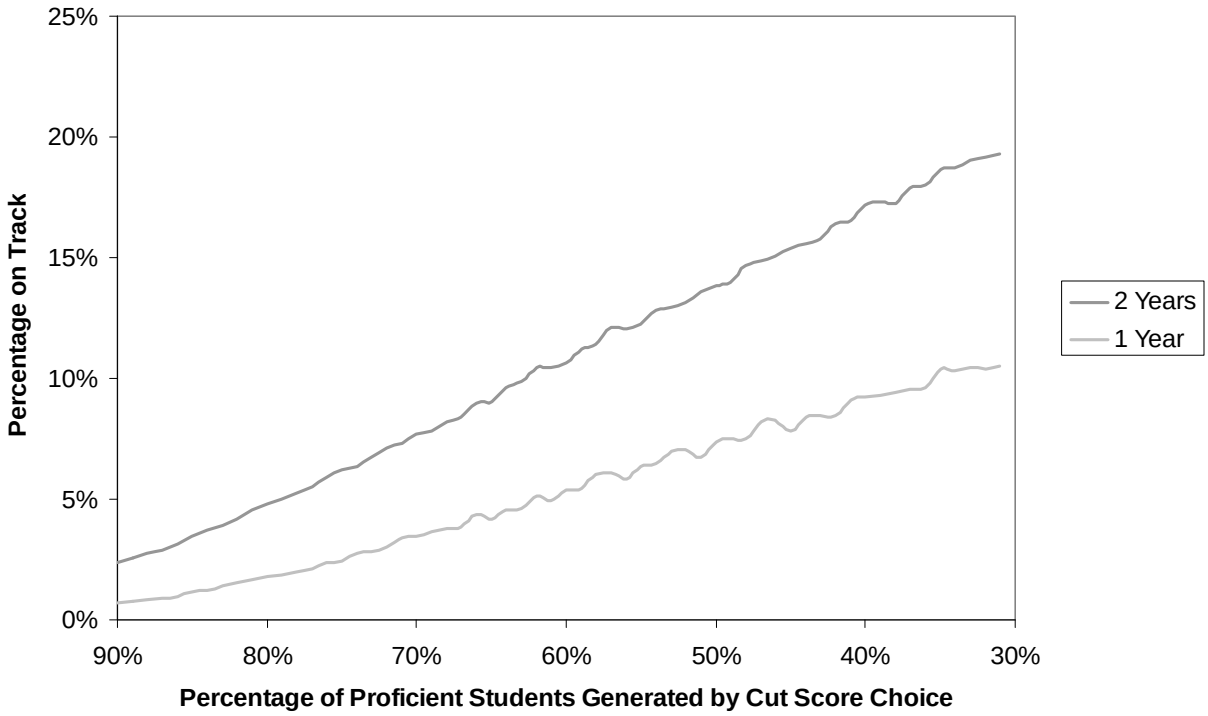

Table 2. Patterns of percentages of proficient students across grades, from decreasing rigor to equal percentages to increasing rigor across grades.

Change	Time 1	Time 2	Time 3	Time 4
+10%pts	45%	55%	65%	75%
+8%pts	48%	56%	64%	72%
+6%pts	51%	57%	63%	69%
+4%pts	54%	58%	62%	66%
+2%pts	57%	59%	61%	63%
0%pts	60%	60%	60%	60%
-2%pts	63%	61%	59%	57%
-4%pts	66%	62%	58%	54%
-6%pts	69%	63%	57%	51%
-8%pts	72%	64%	56%	48%
-10%pts	75%	65%	55%	45%

Figure 4. Theoretical dependence of "on track" students on the articulation of standards as indexed by the change in proficiency rates by grade, ordered from decreasing to increasing rigor.

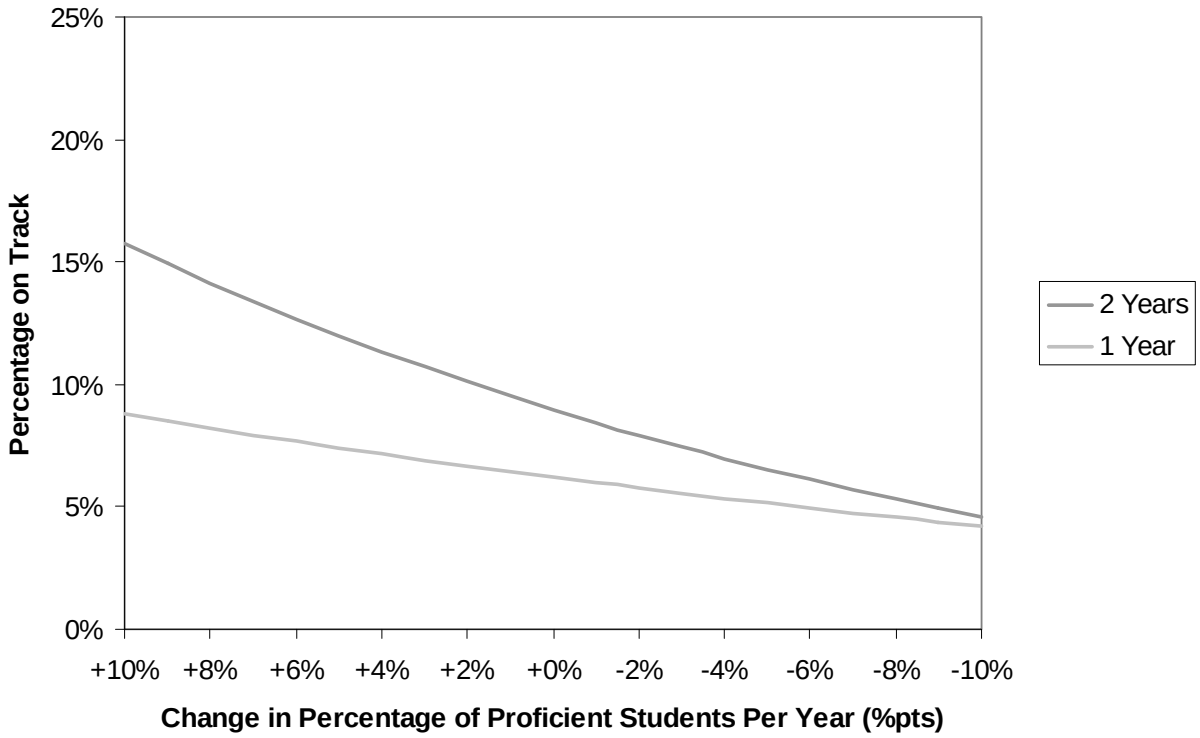

Figure 5. Empirical dependence of "on track" students on the articulation of standards as indexed by the decrease or increase of proficiency rates across grades.

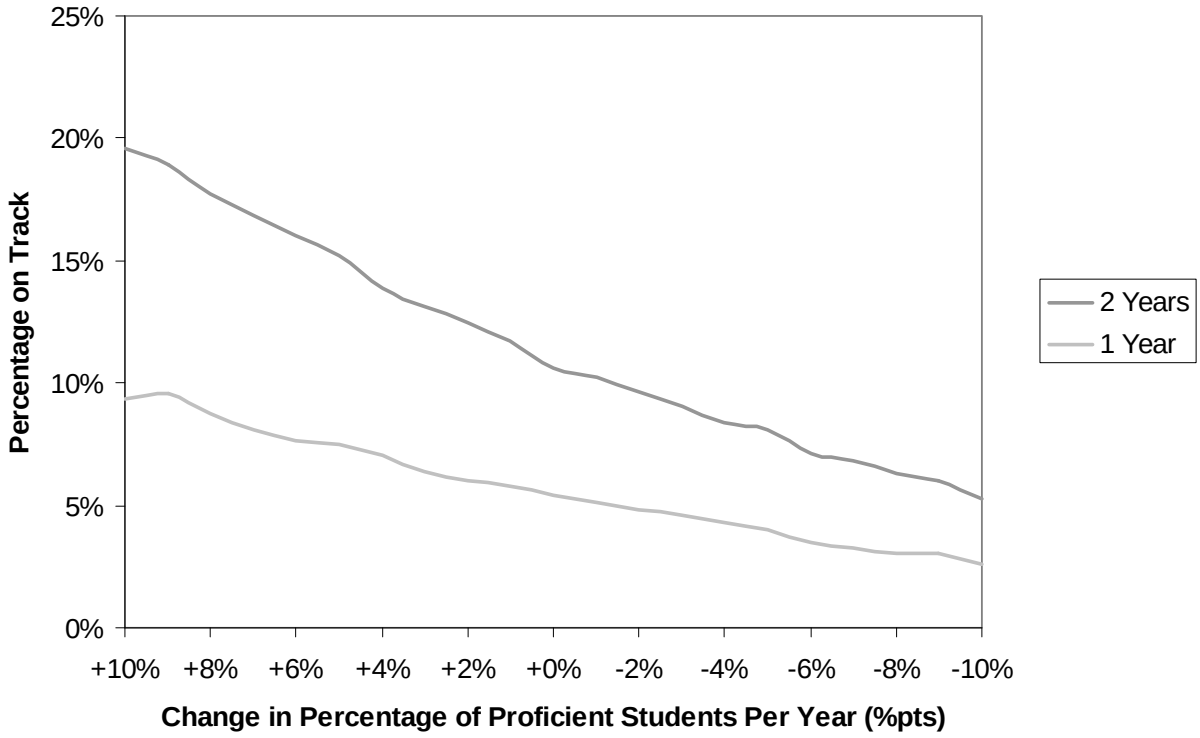

Figure 6. The empirical dependence of the percentage of "growth" schools on cut-score rigor with a set AMO and an AMO that has been raised 5 percentage points. The figure shows results for a time horizon of 2 years and a minimum grade size of 30.

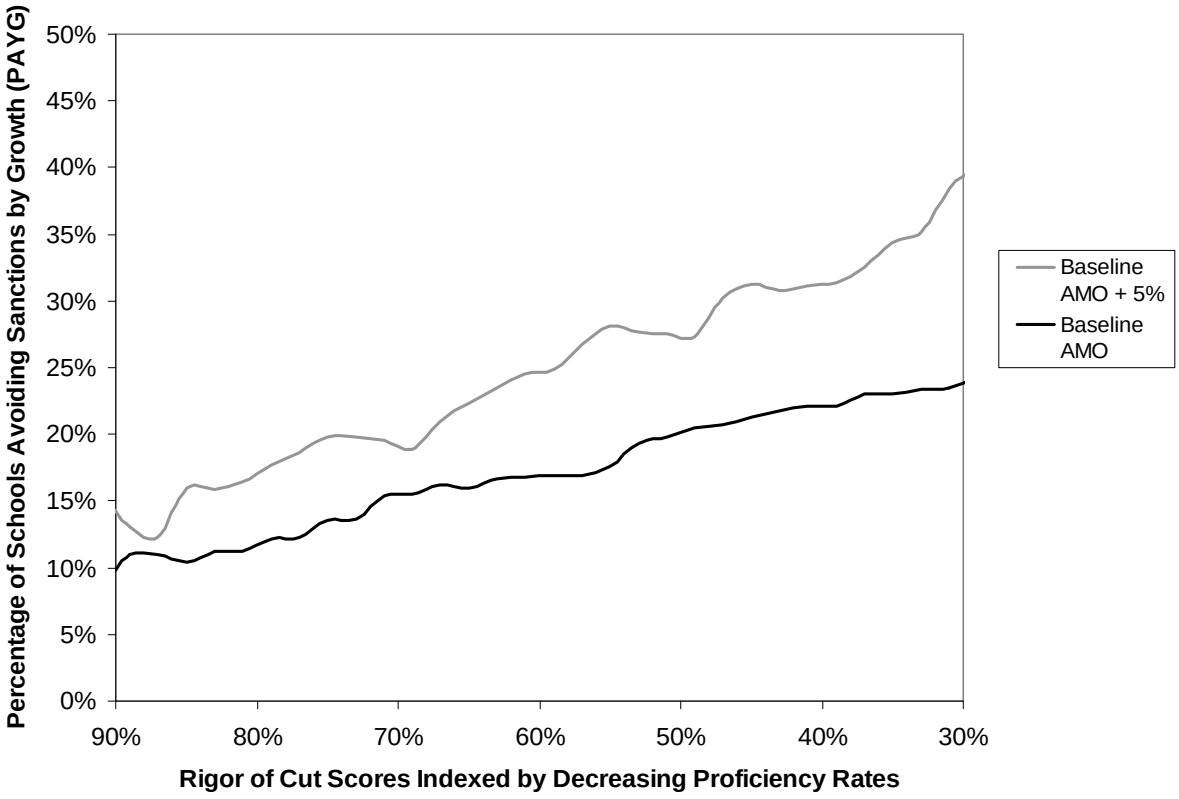

Figure 7. The empirical dependence of the percentage of "growth" schools on cut-score rigor with two fixed AMOs of 50% and 70%. An illustration of a state with 65% proficiency is referenced.