

From No Child Left Behind to
Every Child a Graduate

MEANINGFUL MEASUREMENT

The Role of Assessments in Improving High School Education in the Twenty-First Century

June 2009



© 2009 Alliance for Excellent Education. All rights reserved.

Suggested citation:

L. M. Pinkus, ed., *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* (Washington, DC: Alliance for Excellent Education, 2009).

Ordering information:

Copies of *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* can be downloaded from the Alliance's website at www.all4ed.org. To request print copies of the report, please visit http://www.all4ed.org/publication_material/order_form. The first copy of the report is complimentary. Additional copies are available at a charge of \$1 per copy to cover shipping and handling costs.

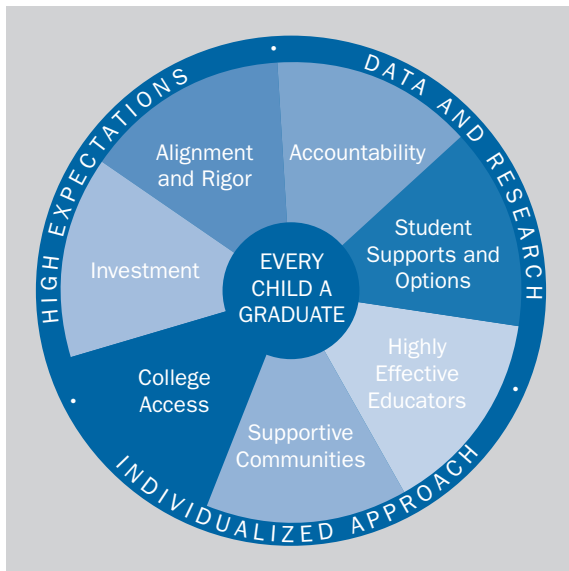
About the Alliance for Excellent Education

The mission of the Alliance for Excellent Education is to promote high school transformation to make it possible for every child to graduate prepared for postsecondary learning and success in life.

The Alliance for Excellent Education is a national policy and advocacy organization, based in Washington, DC, working to improve national and federal policy so that all students can achieve at high academic levels and graduate high school ready for college, careers, and citizenship in the twenty-first century.

The Alliance has developed a “Framework for Action to Improve Secondary Schools” that informs a set of federal policy recommendations based on the growing consensus of researchers, practitioners, and advocates about the challenges and solutions for improving secondary student learning.

The framework, shown graphically here, encompasses seven policy areas that represent key leverage points in ensuring a comprehensive, systematic approach to improving secondary education. The framework also captures



three guiding principles that apply to all of the policy areas. Although the appropriate federal role varies from one issue area to another, they are all critically important to reducing dropouts and increasing college and career readiness.

About the Editor

Lyndsay M. Pinkus is director of strategic initiatives at the Alliance for Excellent Education. Since joining the Alliance in January 2002, she has served in a variety of research, coordination, and advocacy roles, where her work has included managing policy and grant work on a range of issues including graduation rates, data, secondary school accountability, and secondary school improvement, and authoring a number of publications for the Alliance. Prior to rejoining the staff in January 2006, Ms. Pinkus served as a legislative associate at Washington Partners, LLC, providing government relations and policy research and analysis for a variety of clients, including the Alliance. She is a graduate of the School of Public Affairs at American University as a presidential scholar; the Public Affairs and Advocacy Institute at the Center for Congressional and Presidential Studies; and the Institute for Educational Leadership's Education Policy Fellowship program.

Acknowledgments

The Alliance for Excellent Education is greatly appreciative of the authors for sharing their time and expertise in writing the following chapters, as well as of the multiple Alliance staff members and advisors whose dedication contributed significantly to this volume.

The Alliance for Excellent Education is also grateful to Carnegie Corporation of New York for the financial support that made this publication possible.

The views expressed in this volume are those of the authors and do not necessarily represent those of the Alliance for Excellent Education or the funder.

CHAPTER



Measuring Student Achievement Growth at the High School Level

Joseph Martineau

Michigan Department of Education

As education policies at the local, state, and federal levels increasingly include accountability for student achievement, and as the stakes attached to that accountability have risen, interest in various accountability models has grown substantially.

Most accountability models currently in use—including those initially implemented by states to comply with the requirements of the No Child Left Behind Act of 2001 (NCLB)—focus on an absolute level of student achievement. These “status-based” models primarily hold schools, districts, and states accountable for meeting a state-set percentage of students performing at some minimum achievement standard on state-administered assessments.

For example, consider the following scenario of two high schools in a state that uses a status-based accountability model for NCLB. High School A serves a high-challenge student population: only 30 percent of entering freshman scored “proficient” or above on the eighth-grade mathematics

exam. High School B serves a lower-challenge student population: 65 percent of entering freshman scored proficient or above on the eighth-grade mathematics exam. To make Adequate Yearly Progress (AYP) and avoid NCLB-mandated sanctions, both schools need to meet the state-set goal of 70 percent of students scoring proficient or above on the mathematics exam administered to tenth graders. Such status-based accountability models have come under criticism for a number of reasons, including the following:

- Regardless of student background characteristics, risk factors, and incoming education levels, both schools in our hypothetical scenario need to reach the same status target. That means that, over the same period of time, High School A is expected to achieve significantly more progress than High School B.
- There is significant pressure to meet the status goal. As a result, educators in both high schools must focus efforts on meeting the proficiency target, but the impact is greatest in High School A because there is more ground to make up. This has two unintended consequences:
 - First, in both schools, classroom teachers may narrow instruction to focus on the content and test-taking skills that will help students score proficient on the exam, at the expense of other rich content, but the incentive to do so is stronger in High School A.
 - Second, classroom teachers may pay disproportionate attention to the students who are just below the proficient level and can most easily be supported to score proficient and help the school meet its proficiency level. Meanwhile, little attention may be paid to the equally important progress of other students on the performance spectrum, including both those who are furthest behind and those who are already likely to score proficient or above. Again, the incentive is stronger in High School A.

As states have implemented NCLB, criticism of these status-based models has increased, along with calls for a shift to accountability models. This has led to widespread interest in the implementation of “growth models” that value *progress or growth* in addition to *absolute performance*, and that measure—and, therefore, provide incentive to improve—the *progress of all students* along the performance spectrum, not only those *students who perform just below the proficient* bar. Advocates for such accountability models see them as mechanisms for measuring and supporting the goal of improving outcomes for all students over time.

As the policy community looks to the possible inclusion of growth models in the reauthorization of NCLB, there are several issues that need to be better understood. This chapter explains the technical underpinnings of growth models, describes the various types of growth models, states challenges inherent to measuring “growth” at the high school level, and explores implications for policymakers interested in moving toward the widespread use of growth models.

Accountability Models Reflect Expectations

The shift from status-based models to growth-based models would include a significant shift in the balance of expectations within school systems. The expectations implicit in both status- and growth-based models are described below.

Expectations implicit in status-based models

In accountability models based purely on status (such as those currently used by most states to comply with NCLB), the following expectations are implicit:

- All students will be expected to achieve the *same minimum level of achievement* at the same moment in time, regardless of previous level of academic achievement or socioeconomic factors.
- Educators will be held accountable for achieving *different levels of effectiveness* in eliciting student progress depending on the previous level of academic achievement of the children they serve.

Expectations implicit in pure growth-based models

In accountability models based purely on student growth, the following expectations are implicit:

- All educators will be held accountable for the *same level of effectiveness* in eliciting student progress or growth, regardless of the previous level of academic achievement of the children they serve.
- Students will be expected to achieve the *same minimum amount of progress*, regardless of the previous level of academic achievement or socioeconomic factors. In a pure growth model, expecting the same amount of growth from every student regardless of incoming achievement assumes no expectation for closing achievement gaps and no common expectation of ultimate achievement.

These pure models represent the extremes on the accountability spectrum. They also represent the tension between conflicting policy goals: expecting common achievement for all students versus setting common expectations for educators' relative performance. The challenge for policymakers is to implement accountability models that balance the tensions between these goals.

In 2005, the U.S. Department of Education announced a pilot program to allow some states to use an approved growth model for NCLB accountability purposes that counts students “on trajectory toward proficiency within X years.”¹ In establishing the parameters for the pilot, the department defined an appropriate balancing of expectations in this way:

- Educators will be expected to elicit more growth (or learning) in their students whose incoming achievement is below grade level in such a way that on-grade-level competency will be achieved within three years (instead of one).
- Students achieving below grade level will be expected to demonstrate slightly more growth (or learning) than their peers who are achieving at or above grade level.

- Students achieving below grade level will *not* be expected to perform at the same minimum competency level as their peers until three years into the future.

This balance can be further described as delayed but completely common expectations for students, and more similar but not completely common expectations for educators.

One intended major benefit of this prescribed balance is that the achievement or growth of *all* students can count positively toward the accountability determinations of educators, schools, and systems: students already proficient count positively, of course, and any students not yet proficient count positively as long as they are progressing toward proficiency. This is the significant difference from the status models, in which accountability determinations benefit only from extraordinary effort with students already very close to the proficiency target. However, one considerable flaw in this prescribed balance is the unreasonable expectation that educators will effectively move *all* students from below proficient to proficient or above within three years. It can be reasonably argued that students farther from acceptable competency should be given more time to rise to the proficient level, and the amount of time allowed should be based upon an aggressive but commonly observable level of consistent improvement. This approach would make it more plausible that all students can count positively in accountability determinations, by demonstrating that reasonably large numbers of students were able to show the targeted level of growth.

This suggested alternate balance can be described as delayed but eventually common expectations for all students, and similar enough expectations of educators that the expectations are reasonably attainable.

What Are Growth Models?

There are several different types of growth models that can be used to measure growth or progress in student learning, each with different technical requirements.

Types of achievement scales

The foundation for a growth model is the scale on which achievement is measured. The characteristics of the scale define the type of numerical operations that can be performed to calculate growth or progress. There are three important characteristics of scales that have an impact on the types of growth models that are possible to implement: numerical level, span of measurement, and measurement frequency.

Numerical level of achievement scales

Scales on which growth might be measured have typically been described in three broad numerical categories:²

1. **Ratio scales**, in which (a) a true zero exists, (b) the difference between numbers equally distant numerically represent comparable differences in value, and (c) rank order is preserved.

Example: A salary scale, in which (a) \$0 indicates no income, (b) the difference between \$40,000 and \$50,000 represents the same amount of money as the difference between \$1,040,000 and \$1,050,000, and (c) higher numbers always represent greater salaries.

Application to growth in student learning: If a ratio scale is used to calculate student growth or progress in achievement, it is possible to measure that, for example, a student has doubled his previous achievement. However, achievement scales do not in practice have meaningful zero points—what would it mean to have zero mathematics or reading achievement? Therefore, it is unreasonable to expect that such inferences could be made legitimately from a growth model of student achievement.

2. **Interval scales**, in which (a) a true zero does *not* exist, (b) the difference between equally distant numbers represents comparable differences in value, and (c) rank order is preserved.

Example: The Fahrenheit temperature scale, in which (a) 0°F does *not* represent absence of temperature, (b) the difference between 40°F and 50°F represents the same amount of additional heat as the difference between 140°F and 150°F, and (c) higher numbers always represent hotter temperatures.

Application to growth in student learning: If an interval scale is used to calculate student growth in achievement, it is possible to measure that a student has progressed twice as much as a peer, because differences can be compared directly. Many achievement test producers and psychometric scholars claim that an interval achievement scale can be produced, and that therefore differences in growth or progress can be directly compared. However, many other test producers and scholars dispute this claim, indicating that this is only true if the psychometric model used to produce the scales is a true mathematical representation of the relationship between student achievement and answers they give to test questions.* Therefore, it may or may not be reasonable to expect that inferences directly comparing the growth or progress of one student to the growth or progress of another could be made legitimately from a growth model of student achievement.

3. **Ordinal scales**, in which (a) a true zero does *not* exist, (b) the difference between equally distant numbers does not represent comparable differences in value, and (c) rank order is preserved.

Example: Placement in a running event, in which (a) zero does not indicate absence of placement, (b) the difference between rankings 1 and 2 may be minor, but the difference between rankings 3 and 4 may be major, and (c) a higher number always means a longer running time and worse placement.

* The psychometric models typically represent the probability that a student of a certain achievement level will answer a test item correctly. It is clearly reasonable to assume that higher-achieving students generally have a higher probability of answering a test item correctly. However, the exact form of the relationship between student achievement and probability of answering a test item correctly is a matter of debate. Current psychometric models attempt to mirror reality by changing the form of the relationship. The model that best conforms to how things actually happen inside students' heads will produce scores with the best interval-level measurement properties. However, it is impossible to know what the actual form of the relationship should be, because it is unobservable. Therefore, psychometric claims to produce an interval-level scale are unprovable and subjective.

Application to growth in student learning: If the increasing community of skeptics is correct that the existence of interval-level scales cannot be verified, then the scales must be treated as ordinal scales to avoid significant skewing of analyses based on the scales. Achievement scales can be reasonably described as the interval scale at a minimum. Because ordinal scales can be used to compare ranks, it is therefore reasonable to expect that inferences comparing ranks can be legitimately drawn from a growth model of student achievement. If ranks can be compared from year to year, then a growth model is possible but is limited in usefulness by the number of ranks. There are typically four ranks in each grade level, representing far below grade level, below grade level, on grade level, and above grade level in assessments used for accountability under NCLB. To expand the usefulness of growth models, a fourth type of scale is needed:

4. **Ordered interval**, which lies between ordinal and interval and thus is called “ordered interval” here, and in which (a) a true zero *does not* exist, (b) the difference between equally distant numbers represents approximately comparable differences in value for numbers close to each other, and (3) rank order is preserved.

Example: Scale score from an achievement test, in which (a) zero does not mean no achievement, (b) the difference between 100 and 110 is likely approximately equal in value to the difference between 110 and 120, but its comparability to the difference between 310 and 320 is questionable, and (c) larger numbers always represent greater achievement.

Application to growth in student learning: With an ordered interval scale, it is possible to identify certain points on a scale that represent on-grade-level achievement of third graders (say, 300), fourth graders (say, 400), and so on. With such anchor points on the scale, movement along the scale from 300 to 400 represents movement from minimum expected competency in third grade to minimum expected competency in fourth grade, or the target amount of growth from one year to the next to maintain minimum expected

competency. Any movement to measurably above 400 would indicate “more than one year’s growth for one year of instruction,” and any movement to measurably below 400 would indicate “less than one year’s growth for one year of instruction.” However, this only works for students whose proficiency in the first year is at minimum expected competency on the scale.

To expand the usefulness of growth measures, an ordered interval scale can also be used to define additional equivalency anchors. For example, the point of minimum expected competency might be labeled “proficient,” but there might also be another point on the scale above which one would not expect students performing at grade level to score. This point might be labeled “advanced” and identified for third graders (say, 350), fourth graders (say, 450), and so on. For example, with these additional anchor points on the scale, movement along the scale from advanced in third grade to advanced in fourth grade represents moving from above grade level one year to an equivalent point the next. In this case, the amount of growth can be defined as the target amount of growth from one year to the next to maintain above-grade-level achievement, or one year of growth for one year of instruction for such students.

Any number of such equivalency anchors can be defined on an ordered interval scale to expand the usefulness of a growth model in measuring the growth of a student starting out at any point to determine whether a student demonstrated more, equal to, or less than one year’s growth for one year of instruction. For example, one might use the three anchor points on typical state assessments used for NCLB compliance that define four levels, and subdivide the four ranges of the scale based upon the approximate equivalence of differences (or approximately interval-level measurement) in nearby parts of the scale. Such subdivision provides a larger number of ranks that can then be used to measure student growth more precisely than is possible with the relatively wide rankings used for NCLB. For example, rather than having only proficient (300/400) and advanced (350/450) cut score equivalency points for grades three and four, additional equivalency anchors could be added. In this example, one might define equivalency points for grades three and four that identify when a student has reached a middle portion of the proficiency category, and others that identify when a student has reached the high end of the proficiency category.

Span of achievement scales

A second critical characteristic of achievement scales is the span of the scales. There are two types of scale span: multigrade or vertical scales, and separate grade scales. Where possible, it is preferable to create multigrade scales to avoid complications that arise from determining the progress of students when they are measured on different scales from one year to the next.

1. **Multigrade or vertical scales:** In a multigrade or vertical scale, one must place the skills learned in third grade on the same scale as the skills learned in fourth grade, fifth grade, sixth grade, and so on. In some subjects this seems to be a more reasonable assertion than in others. When the nature of the skills changes substantially from grade to grade, it is difficult to claim in good faith that a single scale has been developed.³ For example, in science, grade-to-grade differences in high school subject matter (e.g., earth science, biology, physical science, chemistry) are so stark that it is difficult to claim they can all be placed on a single scale, even though scientific reasoning runs through all four types of content.

In mathematics it might be more reasonable, but one might have to make a break in the scale between the grades where the focus of instruction shifts from basic numeracy to algebraic operations. The introduction of geometry, statistics, trigonometry, and calculus could also be argued to require different scales.

In reading/language arts it might be even more reasonable, but problems still remain. The focus of instruction for younger children might be decoding, sight word recognition, basic comprehension, spelling, and minimally coherent text production. As students move upward in grades, the focus of instruction may shift to fluency, advanced comprehension, rhetorical and grammatical structures, style, voice, and literary criticism. As similar as these skills may be, it is unclear whether they belong on the same scale for younger children as for older children.

Therefore, it may or may not be reasonable to expect that inferences from growth models based on multigrade or vertical scales can

result in valid conclusions about school effectiveness. A more valid approach may be to create separate scales for each grade and address the complexities brought about by measuring student achievement on different scales at different occasions.

2. **Separate grade-level scales:** In separate grade level scales (or nonvertical scales), the only way to measure student progress in achievement is to define anchor points (and bands) of equivalence on the separate scales (as described in the section above on ordered interval scales). This is because in using separate scales, there is no claim made that the scales are measuring the same thing—only that there are points on both scales that can be considered equivalent in evaluating whether students have demonstrated enough achievement. Even with modest changes in content across grades (e.g., in high school science, the scientific method as related to physics versus the scientific method as related to biology), it is possible to identify, for example, that in one year the student achieved far above the acceptable level and in the next year not nearly so far above the acceptable level. With significant changes in content across years (e.g., physics content versus biology content), defining equivalency points on separate scales becomes a tenuous exercise and, therefore, so does the measurement of growth.

Measurement frequency

A final important characteristic of achievement scales is the frequency with which student achievement is measured. Current models are mostly based on annual measurement. However, such infrequent testing is not necessary, and, in fact, measuring different skills at each measurement occasion creates some of the problems discussed above.

The farther apart the measurement occasions are, the more likely it is that the skills being measured will change substantially and qualitatively. If measurement frequency were increased from the typical once-yearly administration, the same scale could be used consistently to measure student growth within a school year or course. For example, the same test*

* “Same test” here means a test that measures the same content and has been placed on the same scale, but does not necessarily contain the same test questions or present them in the same order.

could be used at the beginning and the end of each course, or even more frequently. This is particularly applicable to growth-based accountability models in high school where content tends to change significantly from course to course and grade to grade. The more similar the content across measurement occasions is, the easier it is to measure student growth. While increasing the number of measurement occasions for accountability purposes may be a difficult prospect, it would provide the optimal conditions for measuring growth in high school, where content differs significantly not just from grade to grade, but from course to course.

Types of growth models

Of the many different models that have been termed “growth” models, some measure growth and some measure something else entirely. Each of these models—and the associated measurement scale characteristics required—are described below.

Gain score models simply subtract previous achievement scores from current achievement scores to estimate the amount of growth made by individual students. That individual growth is then aggregated (for a school, district, or state) to estimate the amount of growth observed, on average, in the performance of students served by that school, district, or state. Statistical tests are provided that can differentiate the statistical significance of the growth observed in one group of students from that observed in another. Gain score models require measurement on the same scale at each testing occasion and measurement on an interval or ratio level. This provides a powerful growth model if the assumptions are reasonable. However, such models make suspect assumptions about scale characteristics, and the reliability of the outcomes is limited because gain scores are less reliable than the measurement at either occasion.⁴

Regression growth trajectory models estimate a growth trajectory (or growth rate) for each student, based on each student’s performance on three or more previous measurements. The models may range from relatively simple regression equations to very complex statistical models.⁵ Observed growth rates for individual students are then aggregated for a school, district, or state to estimate the average growth rate observed for students taught in that school or system. Statistical tests are provided that

can differentiate the statistical significance of differences in growth rates from one group of students to another. Regression growth trajectory models require measurement on the same scale at each testing occasion for at least three testing occasions, and require measurement on an interval or ratio scale. Of the different types of growth models, this is the most powerful (if assumptions are met), but makes the strongest demands of measurement scales.

Ordered transition models follow students' rankings from testing occasion to testing occasion (e.g., from proficient last year to advanced this year). The type of transition (which might be classified as some variation on positive, neutral, or negative) made by each individual student is aggregated for an educator, school, or district to describe the types of transition made by students taught in that school or by that educator. These models require measurement data from two or more testing occasions. The models may also range from relatively descriptive models to complex statistical models. Statistical tests may be provided that can differentiate and compare typical transition type across educators, schools, or districts. Ordered transition models do not require either the use of interval-level measurement or measurement on the same scale at each measurement occasion. Ordered transition models require only ordinal or ordered interval measurement.⁶

Prediction deviation models use data about past achievement and/or student and community background characteristics to predict future student achievement and identify the degree to which students underperformed or outperformed their personalized predicted achievement or their personalized predicted growth. These are models that compare expected versus observed achievement and/or growth. In prediction deviation models, the important outcome is how far, on average, a school's or educator's students deviated from what was predicted in terms of either achievement or growth. Statistical tests are provided that can differentiate the degree of positive or negative deviation from expectation from school to school or educator to educator. Prediction deviation models are often called value-added models (or VAMs), because the deviation from prediction is often interpreted as the value added to a student's learning by an individual teacher, school, or district. Prediction deviation models are typically based on any of the previously described types of models, and have the same technical measurement

requirements. A major drawback of this type of model is that any variation in student growth or achievement that cannot be explained by the model is automatically attributed to the school or educator. In other words, the value assigned to individual schools or educators includes the true effects of schools and/or educators lumped together with any sampling, specification, and measurement error in the model.⁷

Prediction deviation models are not growth models, although they are sometimes described as such. They do not qualify as growth models because the outcomes of interest are deviations from expected growth rather than actual student growth.

The “growth” model pilots approved by the U.S. Department of Education are not pure growth models, but **hybrid status/growth models** (also called **on track to proficiency models**). Consistent with the principles laid out by the secretary of education (described above), these models track the progress of not-yet-proficient students toward proficiency within a specified period of time.⁸ Such hybrid models may be based on any of the models previously described, where, rather than simply describing the amount of growth students make in a school or class, the result is whether the student is on track to become proficient within the next X years. (X may be three, four, or five, depending upon the model.) Some of the approved growth models are applied to high schools, as summarized later in this chapter.

Types of Inferences Made from Growth Models

There are two basic types of inferences made from growth models: descriptive; and attributive, causal, or value-added. In a descriptive model, the purpose of the interpretation is simply to describe what has been observed for a given school or system. In an attributive, causal, or value-added model, the purpose of the interpretation is to claim that what has been observed for a given school or system can be attributed solely to the school or system.

There are certain requirements for an attributive, causal, or value-added interpretation to be valid (or at least reasonably approximated). Several researchers, including the author of this chapter,⁹ identify the minimum requirements to be that

- students are randomly assigned to schools and/or educators;
- any missing data must be missing randomly (e.g., students in each school and demographic group are just as likely to miss measurement occasions);
- the technical measurement scale requirements are met; and
- the model contains all appropriate components to isolate the effects of schools/educators.

In reality, it is almost certain that the first requirement is unmet, because students are not deliberately randomly assigned to schools and it is highly unlikely that students are randomly sorted into schools by circumstance. It is almost certain that the second requirement is unmet, because mobility and absenteeism tend to be associated with certain areas or groups of students. The fulfillment of the third requirement can only be logically (rather than empirically) validated, for the reasons described in the section above on numerical level of achievement scales. Finally, the fulfillment of the fourth requirement can only be logically (rather than empirically) evaluated, because one cannot definitively demonstrate that all important factors have been taken into consideration.

Because of these problems with attributive, causal, or value-added interpretations, it is more valid to interpret the results of growth models in a descriptive manner. In practical terms, what this means for accountability models is that schools or systems whose students demonstrate less growth than desired may be called “schools/systems in need of improvement” not because they are poor schools/systems, but because the students they are serving may be in need of schools/systems of extraordinary effectiveness. This is a fine, *but critical*, line demarcating descriptive from attributive (value-added, causal) interpretations.

Summary of Growth Model Requirements

A summary of the minimum technical measurement characteristics required by the different types of growth models is provided in Table 1, followed by qualitative ratings of the statistical and measurement defensibility and validity of each type of model.

Table 1: Summary of Growth Model Requirements

Type of growth model	Minimum technical measurement requirements				Defensibility
	Achievement scale level	Scale span, when measuring		Same scale on all occasions	
		Once per year or less	More than once per year		
Gain score	Interval	Multigrade/vertical	Single grade	Yes	Low to moderate
Regression growth trajectory	Interval	Multigrade/vertical	Single grade	Yes	Very low to low
Ordered transition	Ordinal	Single grade	Single grade	No	Moderate to high
Prediction deviation (not a true growth model)*	Same as base model	Same as base model	Same as base model	Same as base model	Same as base model
Hybrid status/growth (on track to proficiency)*	Same as base model	Same as base model	Same as base model	Same as base model	Same as base model

* Both prediction deviation and hybrid status/growth models can be based on gain score, regression growth trajectory, or ordered transition models.

Growth Model Challenges Unique to High School

High school is the endgame

There is a legitimate argument (as well as a legitimate counterargument) that implementing a growth model in high school may not be appropriate. The argument is that when low-achieving students are in high school, they have very little time to rise to an acceptable level of competency before universal public education has been completed—that high school graduation is the endgame of universal public education, and high schools must be held accountable for eliciting minimally acceptable competency by the time their students graduate.

The counterargument is that high school educators should not be held completely accountable for early education quality. They should be required instead to demonstrate the elicitation of extraordinary student growth

in achievement for low-achieving students—but a reasonably observable ceiling should be placed on that expectation. Some critics believe that accountability for student outcomes should rest more heavily on districts, not individual schools, because for high schools with very low-achieving incoming students it is arguably the district (not the high school) that failed to prepare students adequately for high school education.

The nature of high school subject matter

Because subject matter is much more differentiated in high school than in the early grades (e.g., algebra, geometry, trigonometry, and calculus compared to basic numeracy), it is much more difficult to measure growth. One might ask, “When you say growth in science, do you mean growth in biology, chemistry, physics, earth science, or science reasoning?” The specificity of measurement in high school must also be better differentiated than in lower grades to match the differentiation of content expectations in high school. For example, in elementary science, it may be reasonable to measure growth across years, because the emphasized content across multiple years may be based on the concepts of science in general. In high school, however, where very different and very specific discipline-based content (e.g., physics, biology, chemistry) may be measured from year to year and from course to course, that method would not be appropriate.

The frequency of measurement in high school

In most states, as required by NCLB, student achievement is measured only once in high school for each subject. In most states, this means there is at least a two- or three-year lapse in testing, between measurement in grade eight and measurement in grade ten or eleven. Without multiple years of testing in high school, it is difficult to measure growth. This exacerbates the problems of changes in the nature of skills gauged from measurement occasion to measurement occasion.

Current Use of Growth Models at the High School Level

All of the “growth” models described previously are used for accountability by states approved in the Department of Education pilot. Because of the requirement that students must be “on trajectory to proficiency,” all of the

Table 2: Summary of Growth Models in Place for NCLB

Measurement characteristics			Type of Model							
			Gain score		Regression growth trajectory		Ordered transition		Prediction deviation (VAM)	
Frequency	Level	Scale type ⁴	HS yes	HS no ¹	HS yes	HS no ¹	HS yes	HS no ¹	HS yes	HS no ¹
Once yearly	Ordinal	Single grade						IA ³		
	Ordered interval	Single grade					DE MN ²	MI		
	Interval	Single grade	AK		TX	PA TN			OH ² Typical VAM	Typical VAM
		Multigrade	FL	AZ, AR, MO, NC	CO				Typical VAM	Typical VAM
More than once yearly	Ordinal	Single grade								
	Ordered interval	Single grade								
	Interval	Single grade								
		Multigrade								

¹ Where states have not applied the growth model to AYP it is because of the lack of adjacent grade-level tests in high schools. Generally, where states apply the growth model to AYP, those states also have adjacent grade-level measurement in high schools. They are included in this chart in part to demonstrate that because of the challenges to measuring growth in high school, many states have opted not to include high school in their growth models.

² Ohio and Minnesota apply their growth model to high school AYP even though there is not adjacent grade testing in high schools.

³ Iowa optionally applies its growth model to high school AYP where schools opt to provide grade-ten tests. In this case, growth is followed from tenth to eleventh grade. Where schools opt not to provide grade-ten tests, those schools are excluded from the growth model.

⁴ Some states indicate that they have multigrade and/or interval-level scales, but their growth models do not require such scales. The measurement requirements of the growth models are listed here instead.

department-approved growth models are hybrid growth/status models. The four types of pure growth models—not including prediction deviation—are all represented in the approved growth models, as shown in Table 2 (derived from review of all approved growth model applications, which can be seen at www.ed.gov/admins/lead/account/growthmodel/index.html). Only gain score, regression growth trajectory, and ordered transition models have been

implemented for high school accountability under the growth model pilots. The table identifies the frequency of measurement, the numerical level of measurement scales, and the scale span required by each state's growth model, as well as the type of growth model and whether the state's growth model is applied to high school achievement.

Note that in all cases, the growth models are based upon annual measurement. Note also that only the Ohio growth model resides in the space inhabited by typical value-added models (prediction deviation models). Of the fifteen states with approved growth models, four (Delaware, Iowa, Michigan, and Minnesota) have minimal scale demands: they require only ordinal or ordered interval measurement on scales that do not span multiple grades. Five more states (Alaska, Ohio, Pennsylvania, Tennessee, and Texas) have additional scale demands: they require scales at the interval level, but do not require scales that span multiple grades. Six states (Arizona, Arkansas, Colorado, Florida, Missouri, and North Carolina) go one step further: they require interval-level scales that span multiple grades. Just two of those six states (Colorado and Florida) require a scale that spans grades three through ten rather than just grades three through eight.

There are two other important possible characteristics of a strong growth model for high school. At a minimum, measurement should occur at adjacent grade levels in high school to make the growth model interpretable in terms of individual grades and courses. At a maximum, in order to truly separate out the impact of individual courses on student learning, measurement should occur before and after instruction to measure directly the impact of instruction. This could be at the beginning and end of each course, or as often as at the beginning or end of each unit. Some states have arrived at the minimum by measuring in all high school grades, but none have maximized the usefulness of a high school growth model by implementing pre- and post-instruction measurement. There is one significant caution for a growth model based on pre-/post-measurement—disincentive to encourage students to perform poorly on the pre-test and at maximum competency on the post-test would have to be developed.

Components of an Acceptably Valid High School Growth Model

Measurement components

From a measurement perspective, students should be tested at the very least in adjacent grades so that each individual student's achievement can be tracked from grade to grade. Once-yearly measurement is the minimal measurement requirement needed for a valid high school growth model. If measurement is done any less frequently, it will be impossible to disentangle the growth a student attained in one year's class from the growth that student attained in the next year's class.

For a high school growth model to be useful in making judgments about high school instruction, measurement occasions should occur at the beginning and end of each high school course, to determine how much growth occurred for each student in each class. Because of the specialization in disciplines within a subject (e.g., science, mathematics, language arts) in high school instruction, such a measurement system would allow the growth model to disentangle the growth in mathematics achievement that occurred in, say, an algebra course from growth in mathematics achievement that occurred in a geometry course taken in the same year. This would also enable schools to target professional development efforts in the classrooms where students are making the least progress.

Statistical model

With once-yearly measurement (the minimum measurement requirement), a statistical model would need to take into account qualitative shifts in the types of skills measured in each year by either using a statistical model that does not require interval-level measurement or measuring generic subject matter content rather than subject matter content specific to differentiated disciplines. The disadvantage of the second option is that the information gleaned from the growth model is likely to be less useful in evaluating the impact of specific courses on student growth. The more differentiated the content is from grade to grade, the more difficult it is to validly measure growth from year to year.

With pre- and post-course measurement, a statistical model based on interval-level measurement may be defensible, and would provide strong capacity to differentiate between the student growth or progress observed in one school or class over growth observed in another school or class in which the same course is offered.

Policy Action for Moving Forward with Growth Models for High Schools

The challenges are significant in developing both the necessary assessments and the political will to support a defensible growth model for high schools in every state. Therefore, an aggressive and reasonable target for the large-scale implementation of growth models at the high school level might be eight years. This estimate includes

- approximately two years for the development and passage of federal legislation laying the groundwork for appropriate high school growth models;
- approximately one year for content standards to be developed in accordance with legislation;
- two years for the development of appropriate assessment systems to support appropriately valid measurement of growth in high school;
- two years for the first cohort of students to be measured on the new assessments at least twice, and for pilot growth model analyses to be performed before operational use of high school growth model results for accountability; and
- a final-year application to a second cohort of students, upon whose score data the operational high school growth models would be based.

Requirements for a minimally defensible growth model

Implementing a minimally acceptable and informative growth model will require additional assessments to fill the gap between the last grade of measurement in middle school and the grade level where student achievement is measured in high school that exists in most states. An expansion of testing to these grades would be unpopular, and would require the political will to extend the requirements of NCLB.

Implementing additional tests will necessitate significant additional funding to support the development and implementation of assessments; additional testing-contractor capacity to support development and implementation; and increased capacity of psychometric, statistical, program management, test development, and IT staff for both test contractors and states.

Requirements for a more defensible growth model

Implementing a much more defensible and informative growth model would require either the development of high school course expectations common within each state but different across the country or the development of common high school course expectations across the entire country. There are some states where there are common course expectations, but this is not a universal condition across the fifty states.

The latter is a highly sensitive political issue, as the development of common standards and course expectations is seen by some as a major states' rights issue. However, such a testing system would be much more useful than once-yearly testing in terms of the information that would result from growth models. Such a system would provide disentangled information about student achievement growth occurring in individual courses. In addition, disincentive to encourage students to perform poorly on the pre-test and at maximum competency on the post-test would have to be developed.

The resources needed to implement such new testing programs are similar to those mentioned above, but on a larger scale. Because each course would require a pre- and post-test, the number of tests to be developed in each subject would be twice the number of courses in each subject rather than one test in each grade level where testing is not currently performed. This increase in required resources would even more significantly strain budgets, contractor staff, and state staff.

In order to overcome these obstacles, political will would need to be gathered to pass federal legislation providing additional funding for increased measurement and increased state staffing. Such legislation would also minimally need to require annual testing in every state to fill in the gap between middle school and high school measurement occasions. Finally, in

order to provide a high school growth model that is more than minimally defensible, the legislation would need to require pre- and post-testing in each required high school course, as well as standardization of required high school courses at least within each state, and possibly across the entire nation.

Conclusion

Putting a valid growth model in place for high schools is a worthwhile and important endeavor, in part because measuring student growth is closer to the educational mandate to facilitate student learning than simply measuring students' level of achievement. While growth models do not resolve the tension between setting common expectations for educators and setting common expectations for students, they are capable of balancing that tension in such a way that achievement gaps can be closed without needing educators to perform at unobservable and unreasonable levels.

The challenges to implementing a minimally valid growth model for high schools are significant, and the challenges to implementing an optimally valid growth model for high schools are even more so. In spite of this, however, the emphasis on student learning implicit in growth models, and the usefulness of the information available from growth models for policymakers, educators, and stakeholders, is of sufficient value that the challenges should be taken on.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Joseph Martineau is currently the director of the Office of Educational Assessment and Accountability in the Michigan Department of Education. He received a bachelor's degree in linguistics and master's degree in instructional psychology and technology from Brigham Young University, and a PhD in measurement and quantitative methods in education from Michigan State University. Before assuming his current position, his career included positions as an instructional designer, educational programmer, university instructor, research consultant, psychometrician for the state of Michigan,

and manager of Michigan's K–12 general education assessments. Most importantly, Dr. Martineau and his wife have school-age children directly affected by his work in assessment and accountability.

¹ M. Spellings, "Letter to Chief State School Officers Regarding the Opportunity to Participate in a Growth Model Pilot," <http://www.ed.gov/policy/gen/guid/secletter/080818.html> (accessed on February 21, 2009).

² S. S. Stevens, "Mathematics, Measurement and Psychophysics," in *Handbook of Experimental Psychology*, ed. S. S. Stevens, 1–49 (New York: Wiley, 1951).

³ J. A. Martineau, "The Effects of Construct Shift on Growth and Accountability Models," unpub. diss., Michigan State University, East Lansing, 2004; J. A. Martineau, "Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability," *Journal of Educational and Behavioral Statistics* 31, no. 1 (2006): 35–62; M. D. Reckase, "Controlling the Psychometric Snake: Or, How I Learned to Love Multidimensionality," paper presented at the Annual Meeting of the American Psychological Association, August 1989, New Orleans, LA; M. D. Reckase, "Real World Is More Complicated Than We Would Like," *Journal of Educational and Behavioral Statistics* 29, no. 1 (2004): 117–20; W. H. Schmidt, R. T. Houang, and C. C. McKnight, "Value-Added Research: Right Idea but Wrong Solution?" in *Value Added Models in Education: Theory and Applications*, ed. R. Lissitz, 145–64 (Maple Grove, MN: JAM Press, 2005).

⁴ J. B. Willett, "Some Results on Reliability for the Longitudinal Measurement of Change: Implications for the Design of Studies of Individual Growth," *Educational and Psychological Measurement* 49, no. 3 (1989): 587–602.

⁵ For comparisons, see D. F. McCaffrey, J. R. Lockwood, D. M. Koretz, T. A. Louis, and L. S. Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics* 19, no. 1 (2004): 67–101.

⁶ For examples, see D. W. Betebenner, "Performance Standards in Measures of Educational Effectiveness," paper presented at the 25th Annual Conference on Large-Scale Assessment of the Council of Chief State School Officers, June 2005, San Antonio, TX; R. Hill, "Measuring Student Growth Through Value Tables," paper presented at the 25th Annual Conference on Large-Scale Assessment of the Council of Chief State School Officers, June 2005, San Antonio, TX; J. A. Martineau and D. W. Betebenner, "A Hybrid Value Table/Transition Table Model for Measuring Student Progress," paper presented at the 26th Annual National Conference on Large-Scale Assessment of the Council of Chief State School Officers, 2006, San Francisco, CA.

⁷ D. F. McCaffrey, J. R. Lockwood, L. T. Mariano, and C. Setodji, "Challenges for Value-Added Assessment of Teacher Effects," in Lissitz, ed., *Value Added Models in Education*, 111–41; S. L. Rigney and J. A. Martineau, "NCLB and Growth Models: In Conflict or in Concert?" in *ibid.*, 47–81.

⁸ Spellings, "Letter to Chief State School Officers."

⁹ C. Glymour, "A Review of Recent Work on the Foundations of Cause Inference," in *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, ed. V. McKim and S. Turner (Notre Dame: University of Notre Dame Press, 2007); McCaffrey et al., "Challenges for Value-Added Assessment"; J. A. Martineau, "Distorting Value Added."