

*The Optimal Reference Guide:*

# **Performing on Grade Level and Making a Year's Growth – Muddled Definitions and Expectations Growth Model Series – Part III**

---

*Extraordinary insight™* into today's education information topics

By Glynn D. Ligon



## **ESP Solutions Group**



## Table of Contents

Introduction .....	3
Who Needs Growth Measures Anyway? .....	4
Growth Models that Predict the Future .....	5
Standards versus Norms .....	7
Performing on Grade Level .....	9
Communicating Assessment Results .....	11
Making a Year's Growth .....	13
Can We Agree on a Few Concepts? .....	13
Is it Really Growth? .....	13
Infamous, Briefly Ubiquitous NCEs .....	15
Performance Levels and Growth .....	17
A Year's Growth in a Year's Time .....	18
A Gallery of Illustrations of Longitudinal Performance and Growth .....	19
Current Examples .....	28
Conclusion .....	30



## Introduction

I have a long-standing frustration with two concepts that are too often used in misleading ways. “Performing on grade level” and “making a year’s growth” are used in so many different contexts and with so many different intents that the audiences trying to understand what they mean are too often left with a misunderstanding of reality. Standards-based, criterion-referenced assessments with their results reported in proficiency levels have helped provide some consensus on grade-level performance. However, the frenzy to implement growth models has muddled the meaning of making a year’s gain. In fact, this pendulum-swinging infatuation with reporting growth has knocked us back a few decades and revealed that today’s statisticians haven’t learned the lessons of the past about communicating achievement progress to parents, teachers, and the public.

I apologize in advance for making growth measurement seem murkier than it is already. If you have read the prior two papers on growth models, you already know that I think some statisticians promote methodologies that are overly complex, provide minimal added precision, and are incomprehensible to their audiences. I will label those techniques as **pedantic models** because they ballyhoo sophisticated statistical techniques (e.g., hierarchical linear models, aka HLM) that demonstrate more the statistician’s esoteric mastery of mathematics than an admission of the miniscule, unreliable nature of the differences they tease out of our assessment scores. In other words, the unreliability of data within their models far outweighs the small differences the models squeeze out of the data.

That said, this paper reviews some basic concepts about education assessment scores and how we interpret them. Specifically, two oft-used terms are analyzed to ensure we understand how they are defined.

- Performing on grade level
- Making a year’s growth

The hope is that readers will, agree or not with the conclusions, be well informed when they review and support growth models, and decide how to present the results to decision makers.

## Who Needs Growth Measures Anyway?

Growth is more for the policy makers than for the teachers. The point is that what's happening now with a student in the classroom is what a teacher needs to know. Diagnosing current skills and knowledge is the best indicator of what needs to be done for the student—today. The measurement error and acceleration/deceleration that are endemic to the world of education assessment make the distant past merely interesting, and the projected future academic.

Let me define some of those words within this context.

- **Measurement error** is how far off our assessment score may be from the student's true performance level.
- **Acceleration/deceleration** is the natural changes in the pace of learning that each student demonstrates across the grade levels.
- **Distant past** refers to prior assessment scores that are old enough not to influence a teacher's instructional decisions. The distant past may be less than a year back.
- **Projected future** is the growth model's expectation of where a student will perform in a higher grade level.

Here's that sentence again, now. The measurement error and acceleration/deceleration that are endemic to the world of education assessment make the distant past merely interesting, and the projected future academic. In other words, there is so much error in the measurement and analysis of assessment scores, and students learn faster and slower at different times in their schooling that a student's history of test scores is interesting but not crucial, and a projection of future performance based upon that history intrigues the professors more than the teachers.

## Growth Models that Predict the Future

The definitive growth models are those that report actual growth. Many, however, even those approved by the U.S. Department of Education for measuring adequate yearly progress, attempt to predict future performance. If we required each growth model to illustrate the fully compounded error range around a student's projected score, the science of projections would quickly become thought of as being similar to the Las Vegas point spread for a football game. Sometimes they are right on, more often than not they are off—many times, way off.

Pardon a digression into sports betting. With money on the line, the Las Vegas point spread for college football games is reported to predict accurately the winner less than 60% of the time. On average, the point spread is off by 13 points a game. (These very loose statistics come from scanning the numerous websites that track such arcane topics.)

Again, here are some definitions for those who don't follow college football point spreads.

- Point spread is the difference between the two teams' final scores. (e.g., in 2008, the point spread was 7 with Oklahoma the favorite over Texas—Oklahoma predicted to win by 7 points.)
- 60% means that out of 10 games, the favored team wins by at least the point spread in 6 and wins by fewer points or loses in 4.
- 13 points means that on average, the difference between the final point spread and the Las Vegas line is 13 points. (e.g., in 2008, Texas won 45 to 35, a differential of 17 points from the spread—a fairly typical miss by the odds makers.)

The analogy to predicting student test scores is...

- What if we are accurate only 60% of the time?
- What if our predicted scores are off by an average of 13 points?

With no research to determine reality, these error sizes don't seem unreasonable for growth models predicting over two or three years. States with longitudinal databases should follow the accuracy of the predictions from their growth models. I expect we are only a few years away from having those statistics be available and published by several states.

My prediction: We'll be disappointed in the accuracy of growth models' predictions when we look at them on a student-by-student basis. Consider this. If our predictions are accurate, then what did we learn from them? Only when we are wrong do we have any evidence that our interventions made a difference. The disappointment I am concerned about is from those false negatives—the students we predicted to be proficient, but who were not in the future.



*States with longitudinal databases should follow the accuracy of the predictions from their growth models.*



*Only when we are wrong do we have any evidence that our interventions made a difference.*

Now, let's tie this all back to the definition of a year's growth. When a growth model makes a prediction, it is typically to answer one of two questions.

- Is the student expected to perform at a target level in the future? (e.g., proficient)
- Is the student making a year's growth in a year's time?

In the first question, the required growth to reach a desired performance level varies depending upon where the student most recently performed. Under-performing students must grow at a faster pace in the future than they have in the past to reach a higher performance level.

In the second, the targeted growth should be the same for every student. Wait a minute! Why do we need to make a prediction about this? If our interest is in whether or not a student made progress equivalent to what is typical, then no prediction is needed. There's no need to compound our measurement error and have to explain away all those false positives and false negatives. We merely need to calculate how much the student actually gained and compare that to what we've defined as a year's growth.

Now we are back to the earlier point that predictions are more for policy makers and researchers than for teachers. Do we want to hear teachers saying\*:

- "This kid can coast because the prediction is for future scores to be above the proficiency level."
- "This kid can't make it because the prediction is well below the proficiency level."
- "This kid is not proficient, but the trend says proficient in two years, so everything is fine."

Imagine a fifth-grade teacher grouping students for instruction. Will the teacher want to look at a projection of their performance in two years or a diagnostic measure of where they each perform right now?

---

*\* I have said for decades that the problems we encounter with tests and test scores arise more from misuse than from the nature of the tests themselves. Clearly the examples above are extreme and great effort would be made to ensure those attitudes are not exhibited by teachers. That said, however, how can we be sure?*



## Standards versus Norms

Standards are popular because they establish a goal we want everyone to achieve. Norms are unpopular because they predetermine success for half and failure for half. Instead of arguing these simplistic generalizations, let's explore how they limit our thinking.

- Standards are actually founded in norms. We set standards based upon not only what we want students to know and do, but also what is realistic. The realistic part comes from our understanding of the norm—what students typically know and can do.
- Norms are actually founded in standards. We calculate norms for those standards that are being measured.

When we establish set cut points for standards, politimetrics is used. Politimetrics is the application of psychometrics and politics to make policy decisions about issues such as cut points. A policy body considers normative data before adopting a target such as "the proficient level is a minimum of 70% of the items correct on 70% of the objectives measured." A policy body would not adopt such a rule if they expected only 1% or 99% of the students to meet it.

 **ESP Insight**  
*Politimetrics is the application of psychometrics and politics to make policy decisions about issues such as cut points.*

The universal mediator between standards and norms in assessment is the standard score scale. What a propitious term. The scale score used to report an individual student's precise performance level and to divide students into performance levels (e.g., advanced, proficient, partially proficient, basic, etc.) is based upon a norm-referenced procedure in psychometrics for creating equal intervals between each score. As we all know, equal intervals are required so we can add, subtract, and average scores. Granted, we could just use raw scores (the actual number of items answered correctly), but almost everyone would agree that those have too many limitations.

The fact is that most of the statistical techniques used in growth models are norm-based calculations (based upon normal curves or an actual population distribution). HLM, regression, and other models used for predictions are especially tied to norms, because they rely upon an established relationship between past and future performance that is determined by normative processes (how real students performed in the past).

By the way, this is all good. Standards and norms should work together to provide the most insightful interpretation of academic performance possible.

 **ESP Insight**  
*Standards and norms should work together to provide the most insightful interpretation of academic performance possible.*

Imagine using standards with no norming.

A student answered 75% of the items correctly in grade 4, 79% in grade 5, so all we can project is maybe 83% in grade 6. The grade 6 test may be harder or easier than the others. The proficiency cut point may be higher or lower than in other grades. Typical students may gain more or less than 4 items from grade 5 to 6. We can't look at those factors, because that would be using norms.

For a discussion of norm-referenced and criterion-referenced tests (standards based), see our Optimal Reference Guide, ***Why Eva Baker Doesn't Seem to Understand Accountability, The Politimetrics of Accountability*** (available for free download at [www.espsolutionsgroup.com/resources.php](http://www.espsolutionsgroup.com/resources.php)).

## Performing on Grade Level

In deference to those forward-looking thinkers, we should acknowledge that the concept of a grade level may be evolving. Appropriately, many education agencies and researchers are paying more attention to age-based comparisons for judging academic performance. For now, however, the use of grade-level categories is practical and valid for the great majority of U.S. education systems.

Performing on grade level can be appropriately defined from two very different perspectives.

- **Standards-Based Perspective:** Grade level is defined as the skills and knowledge established as required for a grade level. The boundary for being on grade level is often referred to as the lowest score that classifies a student as proficient.
- **Normative Perspective:** Grade level is defined as the performance level of the typical student in a grade level. Typical, in a normative sense, is the median or 50<sup>th</sup> percentile student; however, a lower percentile may be used to include all students who might have scored at the 50<sup>th</sup> percentile if retested. In other words, on grade level would include those scoring at or above 50 and all others within some unit of SEM (standard error of measurement) or SD (standard deviation) of 50.

From either perspective, the most objective metric for describing a student's status respective to grade level is a score on an assessment. (No need to read into the term assessment whether it is standardized, naturalistic observation, ethnographic, subjective, etc. as long as it produces a score or performance level.) Going back into the psychometric history, we find that grade level performance was defined as the mean/median score for all the students tested at the same time in the same grade level. Yes, yes, this results in half the students being above and half below grade level. That's how it was done. This was straightforward to report and interpret. If more than 50% of the students in a group were "at or above grade level," then that was good.

### Definitions:

**Measure:** the assessment

**Metric:** the scale used to report the score

**Score:** the point on the metric's scale representative of the performance of the student

**Raw Score:** the number of items answered correctly and/or the total points awarded for correct answers

**Grade Equivalent:** the grade level and month of the school year when the average student performs at each raw score

**Percentile:** the percentage of students who score below a raw score (range from 1 to 99)

**Scale Score:** a conversion of a raw score to a scale with predetermined properties such as being equal interval and having a mean of 500 and a standard deviation of 100

**Vertical Scale:** a single, continuous scale score metric that crosses grade levels



*Appropriately, many education agencies and researchers are paying more attention to age-based comparisons for judging academic performance.*

If you remember grade equivalents, then you know that the concept was that a grade equivalent of 5.4 represents the median score of all students tested in the 4<sup>th</sup> month of grade 5. Grade equivalents had some inherent weaknesses that toppled them from their lofty perch back in the early 80's.

- The parents of a fifth grader seeing a grade equivalent of 7.2 wanted to know if their precocious darling should be promoted to the 7<sup>th</sup> grade. (Go ahead folks if you want your star student to suddenly be average.) Imagine what the parent of a 10<sup>th</sup> grader scoring at 18.5 (an attainable score) must have thought.
- The national norms that provided the monthly grade equivalents were established typically only at one or two times of the school year, so the apparent precision of the grade equivalents themselves came from interpolation.
- With the notable exception of the Iowa Tests of Basic Skills (ITBS), grade equivalents were not created as equal interval scales, so adding, subtracting, and averaging them was forbidden.

Despite the shortcomings, we reported grade equivalents from the ITBS when I was a local school district test director (1980-83). Our requirements for the grade equivalent scale were logical and reasonable—and met by the ITBS norms.

- A grade equivalent (GE) associated with the 50<sup>th</sup> percentile must be the grade and month representing the critical norming date (the middle date during the national norming period).
- A 50<sup>th</sup> percentile student must gain 1.0 GE a year to maintain the 50<sup>th</sup> percentile ranking.
- Students above the 50<sup>th</sup> percentile must gain more than 1.0 GE to maintain the same percentile.
- Students below the 50<sup>th</sup> percentile must gain less than 1.0 GE to maintain the same percentile.

Percentiles became the most popular metric used for reporting test scores. Easy to explain, but percentiles are not without their own troubles.

- Percentiles cannot be averaged. (Not equal interval.)
- Users at times think a percentile represents the percent of items answered correctly.
- Some parents think a 65<sup>th</sup> percentile is a failing grade.
- A student at the 50<sup>th</sup> percentile two years in a row may be thought not to have grown any.
- A student at the 25<sup>th</sup> percentile two years in a row may be thought to have kept pace with the average students.
- A student moving from the 12<sup>th</sup> to the 14<sup>th</sup> percentile may be thought to be catching up with the average student.
- A student scoring at the 65<sup>th</sup> percentile in both reading and mathematics may be thought to be equally above grade level in both areas.

All this tells us that for measuring and reporting growth, grade equivalents (unless they are equal interval), percentiles, and raw scores have shortcomings. Scale scores work—if they are equal interval. If scale scores are also a vertical scale, even better.

## Communicating Assessment Results


I learned my practical psychometrics and how to report and explain test scores in a community full of university professors, graduate students, native Spanish speakers, politicians, news media, poverty-level families, and educators. Precision in communicating results was an imperative. Sloppiness in reporting was never an option.

H. D. Hoover, ITBS author, University of Iowa professor, was in Austin (TX) one evening when I was scheduled to present annual test results to parents at an NAACP meeting. Seems Iowa City at the time had not offered H.D. such an opportunity, so he joined me at the meeting. He was great. The parents were great. The lasting insight I gained from that evening is that averages don't mean much to parents. They generally already know how the averages are going to look. The black parents at the meeting already knew that on average their students scored below the white students in Austin. Beyond knowing where their own children performed, they wanted to know about individual students who were exceptions to the average. Did some students in their community score at the highest levels of the test? Yes. Did some students grow impressively and keep up with the highest performers in the city? Yes. An average may have shown students from their schools performing poorly, but seeing that some students were above grade level and growing at an impressive pace was encouraging.

From another perspective, Dr. Evangelina Mangino and I conducted an annual, call-in television broadcast at the time test scores went home to parents. Predictably every year, parents would call in to "complain" that their children scored 99<sup>th</sup> percentile every year, so they either aren't learning anything, or the tests were inadequate. Clearly, percentiles were not adequate to describe growth for these students and parents. Grade equivalents helped, but as described elsewhere, created their own problems.

Now to the point of all this. We must know how to communicate with all audiences when we report assessment results—for groups or individuals. Don't make the mistake of thinking that reporting for individuals and groups is the same. When we report for an individual, measurement error is the key to the confidence we should place in the score. When we report averages for groups, sampling error is the key. However, when we report counts of students in performance levels, measurement error is again the key. For a full discussion of these issues, see **Confidentiality and Reliability Rules for Reporting Education Data** (available for free download at [www.espsolutionsgroup.com/resources.php](http://www.espsolutionsgroup.com/resources.php)).

From the curriculum standards perspective, performing on grade level means mastering the skills and knowledge adopted for a grade level. Many papers have been written about the relative rigor of these standards across subject areas, grade levels, and states, but the bottom line is that the definition of on-grade-level performance is typically clear.

 **ESP Insight**  
*We must know how to  
communicate with all  
audiences when we report  
assessment results—for  
groups or individuals.*

With the emergence of criterion-referenced assessments and the impetus from the No Child Left Behind Act (NCLB), performing on grade level has defaulted to being proficient or above on the state assessment.

This is great. Now we can all agree that performing on grade level means that a student is at or above some established performance level on an assessment.

Whether or not that performance level is the traditional 50<sup>th</sup> percentile, the omnipresent proficient level, or a less rigorous partially proficient status, there is a specified point on the score distribution that defines the bottom limit for grade level performance.

So what we must insist upon whenever we hear a report of students who are performing on grade level is that the definition of that venerable status be clearly described.

## Making a Year's Growth

### Can We Agree on a Few Concepts?

- If a student falls farther below grade level from one year to the next, that student could not have made a year's growth.
- If a student rises farther above grade level from one year to the next, that student must have made more than one year's growth.
- Even the students who fall farther behind grade level each year are making some growth.
- For a student to remain highly ranked among peers, that student must make more than a year's growth every year.

In other words, the persistent high achievers continue to grow faster and farther above grade level annually. The persistent low achievers continue to grow slower and become farther below grade level each year.

With these global assumptions or agreements, we can move on to define a year's growth.

### Is it Really Growth?

Let's talk about artifactual growth for awhile. This is important because not all growth is good enough, not all growth is as good as it's purported to be.

We want our growth designations to be real, valid, and understandable, not artifacts of the model used. Here's a definition of an artifact from the medical community.

*The American Heritage® Medical Dictionary Copyright © 2007, 2004 by Houghton Mifflin Company. Published by Houghton Mifflin Company. All rights reserved.*

#### **artifact, n**

1. anything made by human hands or activities.
2. a product that may develop during an analysis performed to identify the composition of a substance. Mainly a consequence of the conditions of the analysis.

**Artifacts** are misrepresentations of tissue structures seen in medical images produced by modalities such as Ultrasonography, X-ray Computed Tomography, and Magnetic Resonance Imaging. These artifacts are caused by a variety of mechanisms, such as:

- The underlying physics of the energy-tissue interaction (i.e., Ultrasound-air)
- Data acquisition errors (mostly from patient motion)
- A reconstruction algorithm's inability to represent the anatomy

Before we struggle with defining a year's growth in a year's time, let's throw some light on how growth may be inappropriately defined.

**Artifactual Growth:** Growth that is misrepresentative of reality because it is in relationship to a false standard.



*The persistent high achievers continue to grow faster and farther above grade level annually. The persistent low achievers continue to grow slower and become farther below grade level each year.*



**Artifactual Growth:** Growth that is misrepresentative of reality because it is in relationship to a false standard.

Artifactual growth is present when the determination of success or failure is an artifact of how the growth is measured or reported rather than the significance of the growth itself. Examples of artifactual growth include:

- Value-add growth measures that adjust for demographic, programmatic, and prior performance factors such that students who gain less than a year's growth are characterized as successful. To be fair, they are more successful compared to similar students.
- Student growth percentiles (SGP) above 49 for low performers. SGPs are defined and discussed later.
- Maintaining a proficient performance level, but falling significantly within that level.

In each of these cases, the appearance of positive growth is an artifact of how growth is measured and reported.

I must say I am frustrated with the loose way we all talk about students' achieving a year's academic growth. The frenzy to find a beneficial growth or value-add model to validate school effectiveness has added to the confusion. Actually, I would be pleased if people were confused. The reality is that almost all of us think we know what a year's growth means. A bit more outward expression of uncertainty would be encouraging.

So one goal of this paper is to confuse, to shake people out of the complacency that surrounds our interpretation of student proficiency measures in this era of fascination with growth.

Simply put, if your education agency is reporting how many students are making a year's growth in a year's time, then that report may be misleading. If, as a parent, you receive a report that your child has made a year's growth, look farther into what they really mean.

As Congress and a new Secretary of Education revisit the No Child Left Behind Act of 2001, they need to beware defining artifactual growth as acceptable and successful. Ironically, those high-performing schools that were torpedoed by NCLB's multiple indicators and subgroups have a stake in this as well. The artifactual growth of some methodologies under-represent the size of the gains made by high-performing students.




## Infamous, Briefly Ubiquitous NCEs

In the final era of the Elementary and Secondary Education Act (originally enacted in 1965; later called the Improving America's Schools Act in 1994; then called in 2001 the No Child Left Behind Act), Chapter 1 was substituted for Title 1 as the compensatory program name and, of even more interest here, normal curve equivalents (NCEs) were instituted to express the growth of students on tests (now called assessments). NCEs are a simple concept. Percentile ranks were normalized, converted to z scores with a mean of 50, standard deviation of approximately 21, and a range from 1 to 99. This took the flat percentile distribution and created an equal-interval scale with scores that could be added, subtracted, and averaged. (At least they could be if one were to apply loosely the underlying assumptions for the data that created the original percentiles.)

**Disassumptive Analyses** are statistical analyses that employ a model with which the data being analyzed do not meet the assumptions or requirements for the data. The analyses are disassumptive because they violate the foundational assumptions that determine when their use is valid.

Example 1: If a report were to provide mean percentiles, that would violate the assumption that data that are averaged must be equal interval. Percentiles are merely ordinal.

Example 2: If a growth model were to use HLM to predict scores from only past scores without using data on how real students perform on assessments in higher grade levels, that would assume that the growth curve is the same across all grade levels, which is inconsistent with the actual data.

 **ESP Insight**  
*Disassumptive Analyses are statistical analyses that employ a model with which the data being analyzed do not meet the assumptions or requirements for the data.*

The requirement for Chapter 1 was to report annually what percentage of students maintained or improved their NCE score. The presumptive assertion was that this defined students who were making an acceptable growth on the assessment. Kudos go to Chapter 1 for establishing a nationally standardized way to report growth for accountability. There were, however, several logical cracks in this methodology.

- Chapter 1 growth was compared to growth by non-Chapter 1 students by projecting out the growth line for Chapter 1 students and seeing if the non-Chapter 1 student line was on a higher or lower trajectory. Guess what—higher.
- Chapter 1 students were highly mobile, so the requirement to have NCE growth measures for 60% of the students served was difficult to meet.
- Successful Chapter 1 students were exited from service, so Chapter 1 programs were constantly replacing their successes with more challenging new low-performers.
- The double whammy was that those successes moved into the comparison group of non-Chapter 1 students.
- Because Chapter 1 students were the lowest achievers in a school, they typically scored well below the 50<sup>th</sup> NCE; therefore, maintaining or even

slightly improving that NCE the next year did not equate to closing the gap or even keeping up with average students.

I'm not sure whether or not Chapter 1 really equated maintaining one's NCE from one year to the next to a year's growth. However, while equal NCE gains from one year to the next for both high and low achievers were theoretically equivalent, maintaining the same NCE from one year to the next required more learning the higher up the scale a student performed.

During the time NCEs were required for Chapter 1, test publishers offered them along with percentiles in reports. NCEs are still found as a psychometric anachronism in some assessment reports or program evaluations.

## Performance Levels and Growth

Two quick illustrations:

- A non-proficient student outperforms other non-proficient students from one year to the next—a year's growth? Possibly not.
- An advanced student maintains the same relatively high-performance level from one year to the next—a year's growth? Much more than one.

I will admit that at times I can split hairs, but in this case, the splitting is important. Important is defined as “we are misleading parents and teachers about how well their students are performing.” We are overstating how well some low performers are growing and we are understating how well some of our high achievers are performing.

When we report group statistics, this blurring of the precision of growth is understandable. But, when we report individual student gains, precision is a requirement.

One of the admirable aspects advanced by NCLB is the counting of every student rather than the averaging of all students' performance.

Is this a new issue? Certainly not. In fact, when I was a local school district test director back in the 80's, we were struggling with the same definitions and reporting challenges. What's changed today is that there seems to be pedantic thinking on the part of some experts who are guiding educators in how to measure student performance and report the results. To be kinder, the problem is most likely a case of textbook formulas and terms being applied to education assessment data without full understanding of the context of the education environment. In other words, some experts know their mathematics and statistics much more than they know schools and students.

One characteristic of the methods I prefer is simplicity and, when simplicity is unattainable, then transparency. There should be no black boxes. If not yourself, someone you trust must be able to replicate the calculations of any growth model under consideration.



*We are overstating how well some low performers are growing and we are understating how well some of our high achievers are performing.*



*Some experts know their mathematics and statistics much more than they know schools and students.*

## A Year's Growth in a Year's Time

For the growth advocates, the ultimate benchmark is making one year's growth in one year's time. This standard means that a student has made the amount of progress that has been officially adopted as representing what the student should have learned in the prior grade level. Clearly more is learned by many students, but this benchmark is the gold standard for judging every student. Unfortunately, this standard means so many different things to different people.

Using all we have discussed above, there are two definitions of a year's growth that emerge. A distinction between a standards-based definition and a norm-referenced definition is again useful.

Making a Year's Growth:

- **Standards-Based Perspective:** Maintaining or improving the proficiency level from one year's administration to the next (Maintaining may only apply to students at the proficient level or higher.)
- **Normative Perspective:** Making a scale score gain from one year to the next that is equal to or greater than that made by a 50<sup>th</sup> percentile student

Assumptions for These Definitions:

- A large-scale assessment is conducted at multiple grade levels, preferably consecutive grade levels.
- The assessment produces scale scores that are equal interval within each grade level. (Vertical scaling is not assumed.)
- The assessment produces percentile rankings aligned with these scale scores. (State-level percentiles are assumed for state-level definitions of growth.)
- The assessment is standards-based whether or not its psychometrics would categorize it as criterion-referenced, norm-referenced, or both.
- Individual student scores are linkable across administrations.

Already, I can imagine you questioning these definitions. That's what the rest of this paper does as well. In the end, the basic concepts underlying these definitions will have been illustrated to support these definitions.

Some definitions of a year's growth that are clearly wrong include:

- Maintaining the same percentile level from one grade to the next (except for the 50<sup>th</sup> percentile)
- Making the same growth as other students with the same prior score (except for the 50<sup>th</sup> percentile)

Some may argue that a year's growth is relative to the prior achievement level of a student. Wrong. When a reasonable person talks about a year's growth, that person is thinking of growth for an average student. Yes, a low achiever can make growth equivalent to that of other low achievers, but try to defend that as being a full-year's growth when reporting assessment results to the public—or to that student's parents.

## A Gallery of Illustrations of Longitudinal Performance and Growth

For awhile, forget about the extreme outliers—those progenies with severe disabilities or prodigies with rare talents. Statistical analyses work so much better if the outliers are somewhat ignored, and the remaining 97% of our students who meet the assumptions of the mainstream assessment measure are included. That's another paper topic—how to include exceptional students when administering and reporting the results from large-scale assessments.

The following series of graphs illustrates basic concepts that are foundational to a rationale for measuring and reporting academic growth as measured by assessments.



*Statistical analyses work so much better if the outliers are somewhat ignored, and the remaining 97% of our students who meet the assumptions of the mainstream assessment measure are included.*

The eight observations illustrated are:

1. **Growth plateaus.** Growth as measured by assessments is greater in the early grades.
2. **Variance increases.** As the grade levels rise, the scores made by students on assessments spread out and the range across them increases.
3. **50 represents.** The 50<sup>th</sup> percentile is an important reference statistic for interpreting individual student performance at each grade level.
4. **Error influences.** We can place too much confidence in test scores that are not as precise as we'd like.
5. **Standards lag.** The standards measured at each grade level tend to reflect more standards from earlier grade levels as students progress through high school.
6. **Projections soar.** Projections made from elementary grade assessment performance tend to over-estimate performance in higher grade levels.
7. **Baselines rule.** Establishing a baseline for comparison of future performance allows progress to be determined.
8. **Students diverge.** As grade levels rise, the gap between low and high achievers widens. A high achiever must demonstrate greater growth each year to maintain that gap. Typical growth for a low achiever allows the gap to increase.

Real growth can be any one of these.

1. Making more than a year's growth in a year's time. (A sign of success for low achievers, but not necessarily for high achievers.)
2. Growing enough to improve a low performance level or maintain a high performance level. (Remaining proficient or advanced; or moving up to proficient or advanced.)

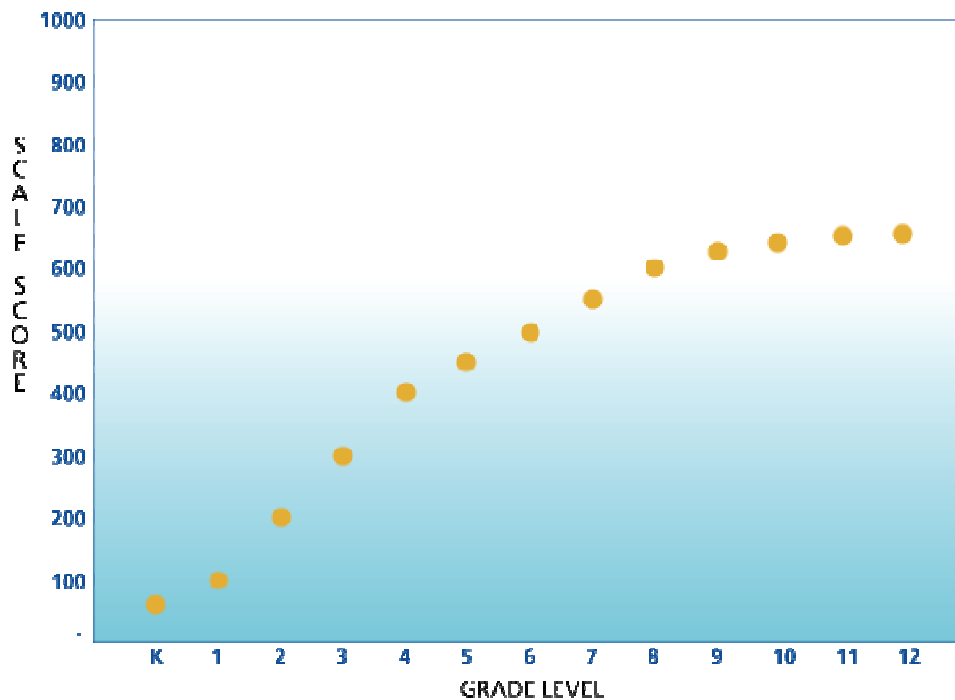
Artifactual growth can be any one of these. (Artifactual growth is further illustrated by Colorado's student growth percentile model, which is discussed later.)

1. Growth under-represented—Growing less than other high achievers. (Still growing more than an average student or more than a year's growth.)
2. Growth over-represented—Growing more than other low achievers. (Still growing less than an average student or less than a year's growth.)

### Observation 1. Growth plateaus.

- Examining numerous quasi-longitudinal achievement graphs over the decades has produced a consistent observation. Students make the greatest gains on assessments in the early grades. Several factors might account for this phenomenon.
  - Standards and what is to be tested are easier to define and differentiate into their components for the early grade levels. Thus, psychometricians can create measurements that are more sensitive to teachable concepts that students can learn efficiently.
  - Secondary coursework is content and skills oriented with students taking a wider variety of offerings that extend beyond the scope of the assessments.
- Growth as measured by assessments is almost flat in high schools. There may be considerable learning taking place above the ceiling or beyond the scope of the assessment.

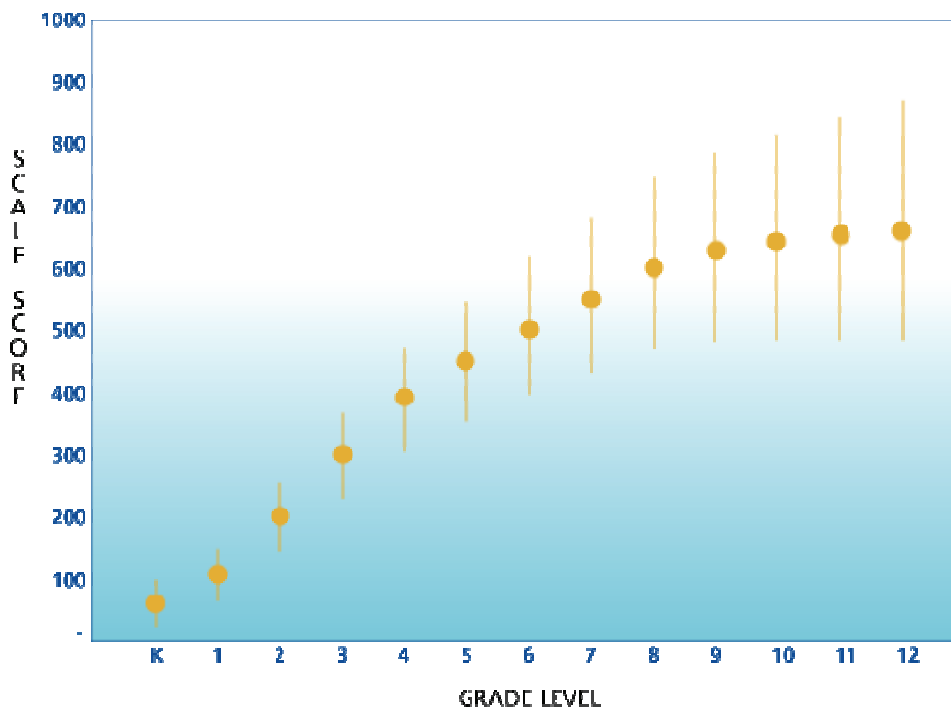
*Why is this important? Growth models that use only past scores for students typically predict higher future performance than models that use real scores for higher grade levels.*



### Observation 2. Variance increases.

- The variance or spread of assessment scores increases as the grade level of students tested rises. In other words, students vary more in their performance levels as they get older.
- The content and skills of assessments appear to expand to greater ranges as the grade levels rise.
- As grade-level assessments advance in their difficulty, there is more room for individual students to differ in their performance.

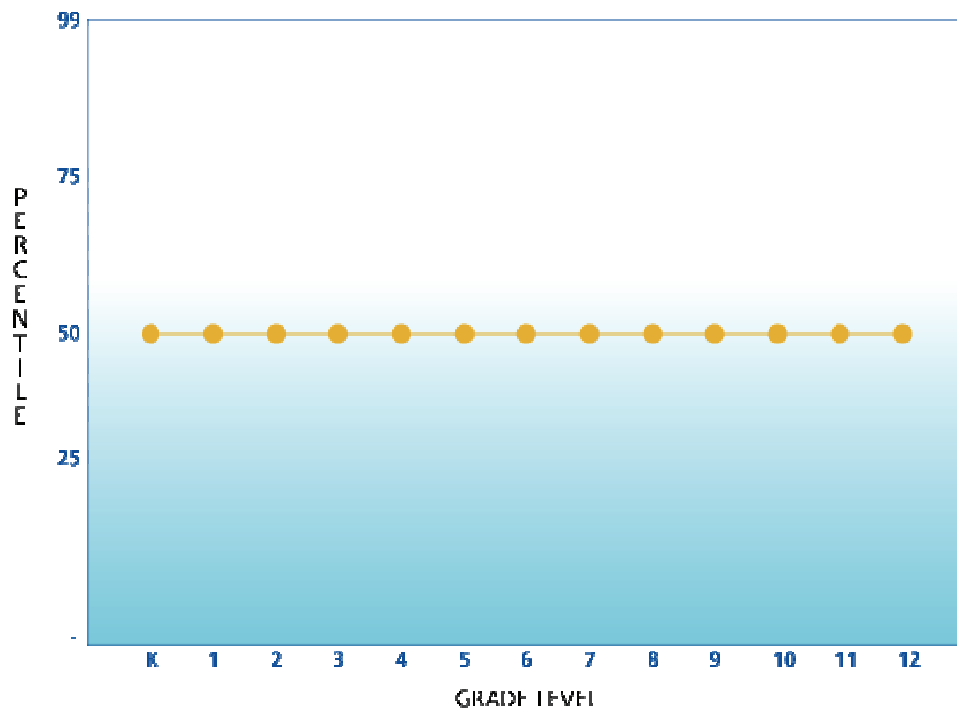
*Why is this important? Growth models that predict performance are looking into a future where the variance, in this case the error, of those predictions gets higher with each grade level.*



### Observation 3. 50 represents.

- There may not be an actual average student who remains average throughout an entire school career, but the hypothetical average student would score at the 50<sup>th</sup> percentile every year.
- The composition of the student population changes from grade level to grade level. Not all students attend public kindergarten. Many who attend private schools join the public schools in middle and high school. However, in high school, dropouts begin both reducing the enrollment totals and removing from the population some of the lowest performers.

*Why is this important? The average student at each grade level provides a context, a reference, a benchmark that is useful for interpreting academic performance.*

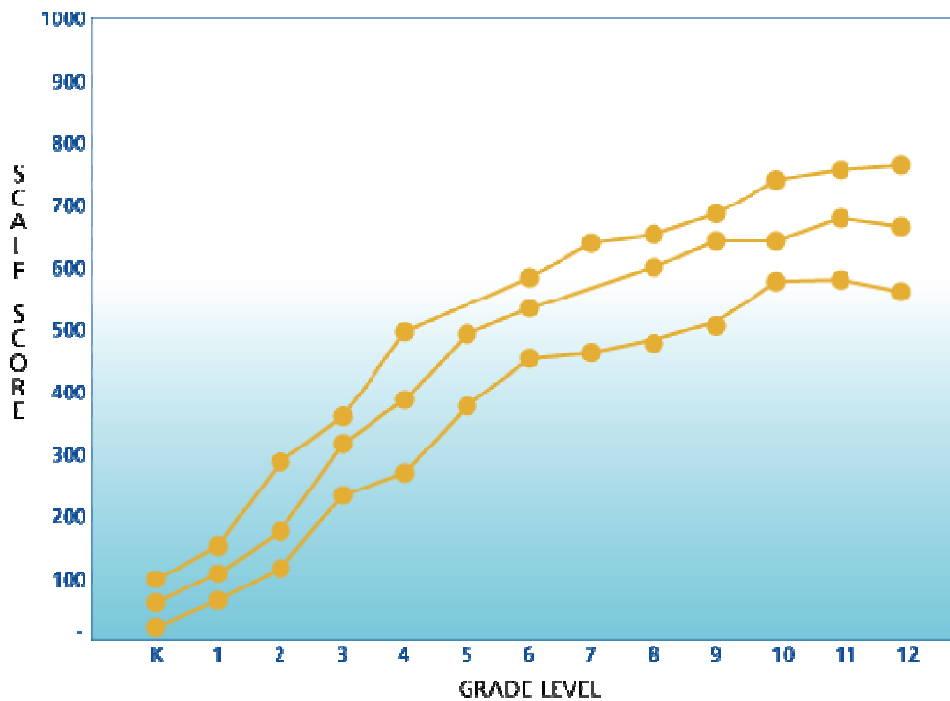




#### Observation 4. Error influences.

- Students wobble their way from one grade level to the next with learning fits, spurts, and plateaus.
- Measurement error also adds to the lack of precision in our test scores.
- Variance (standard deviation) reflects this wobble and error, and results in our becoming less precise in our measurements.

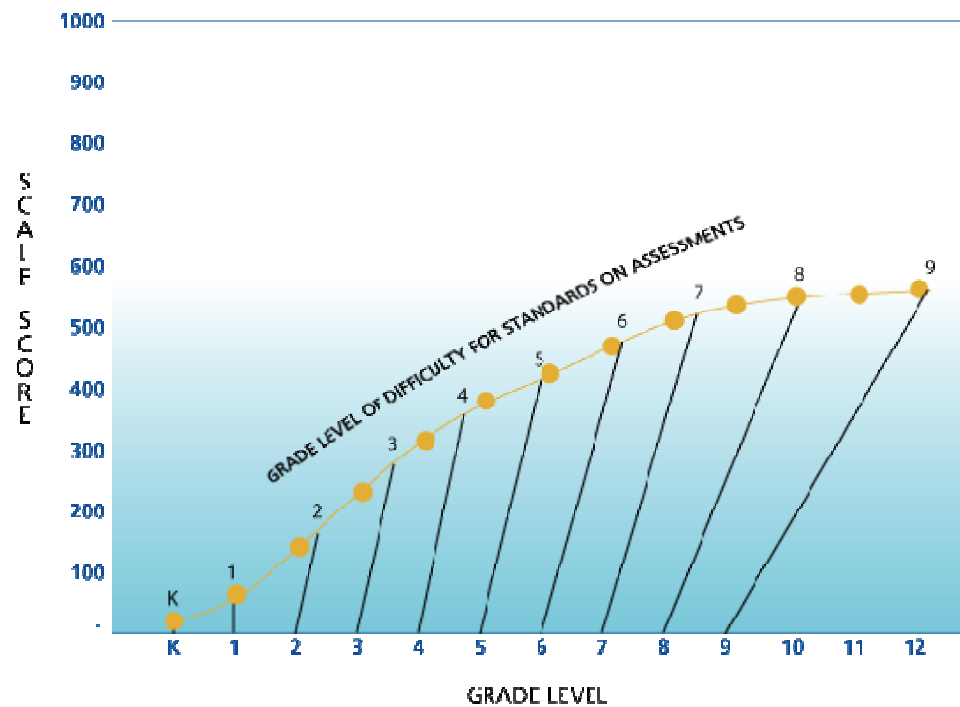
*Why is this important? We probably place too much confidence in one year's test score—definitely too much confidence in a growth measure based upon more than one score.*



### Observation 5. Standards lag.

- The standards that are measured on assessments fan out, span a larger range of grade levels as the students get older.
- The typical standard being measured at the end of high school may be tougher, but the rate of the rise in the difficulty level of the test items slows down.
- Standards measured at the 11<sup>th</sup> and 12<sup>th</sup> grade levels on state assessments may be equivalent only to the average performance measured at the 9<sup>th</sup> grade level.

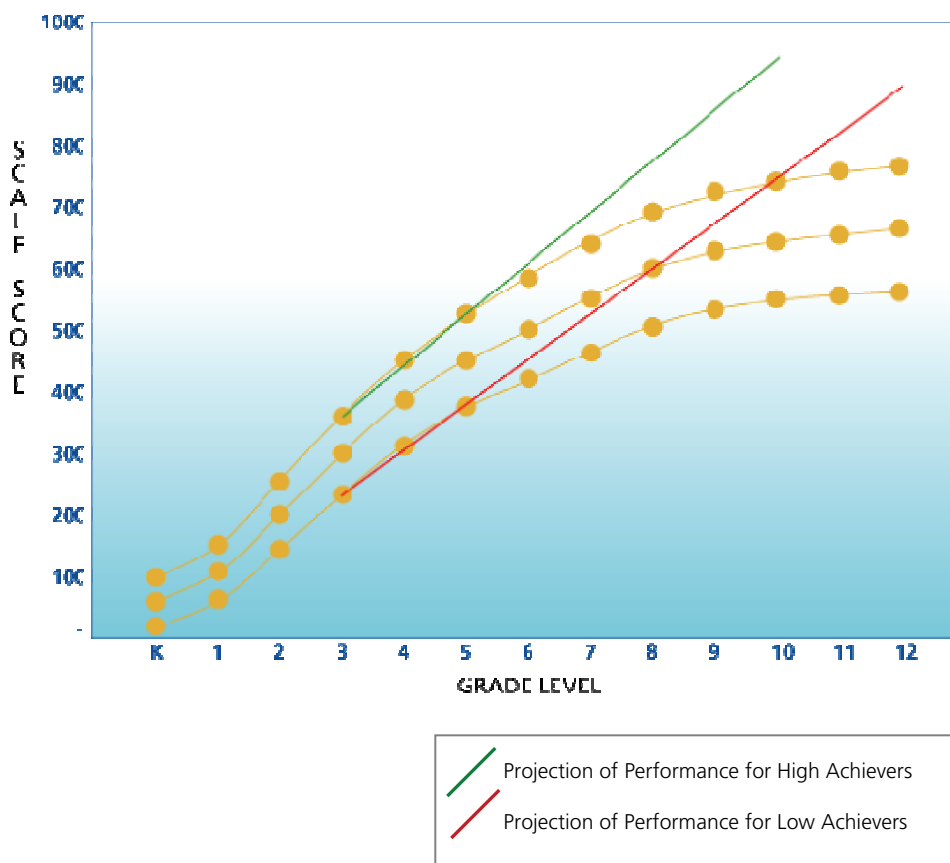
*Why is this important? When we interpret performance on tests, we should understand that being on grade level or being proficient is less of a rigorous standard in secondary schools.*



#### Observation 6. Projections soar.

- When we use only past performance in elementary grades to project future performance, we can launch projections that give us a false sense of confidence in how well students will perform in the future.
- High achievers can be projected to soar above the ceiling of the assessment, while low achievers can be projected to rise to the top.
- A curvilinear model is needed to capture the change in pace of performance across grade levels.

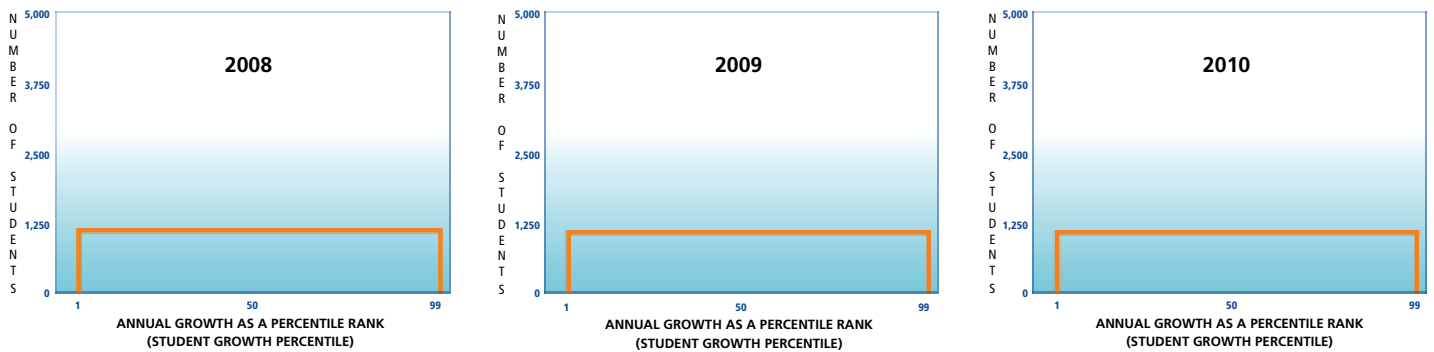
*Why is this important? We mislead ourselves and others when we project future performance that is too high. At-risk students who need extra support and intervention are falsely classified as safe.*



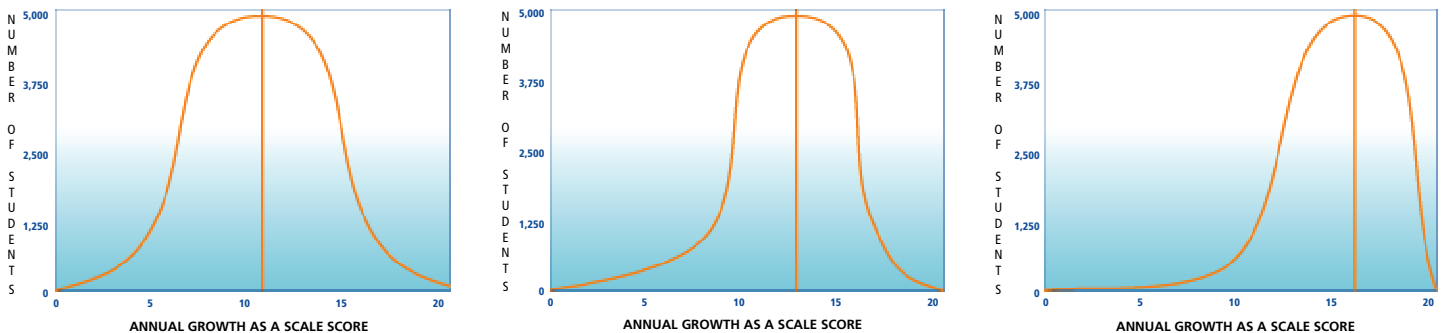
### Observation 7. Baselines rule.

- If a true normative methodology is followed, about half of the students would be above and below average in their growth each year.
- Establishing a baseline year for comparison allows every student to exceed or miss that baseline in future years.
- Reporting growth as a percentile rank results in a fixed number/percent of students at each growth level each year.
- An antidote to this restriction is setting the percentile ranks associated with each raw score point in a baseline year rather than reporting percentiles calculated annually for each new cohort of students.
- Reporting growth using a scale score (one founded in a baseline year) allows as many students as can to make growth greater than a year's equivalent (or less if schools are not effective).

*Why is this important? Adoption of a growth model should not result in a limitation of the number of students who can be successful.*



A snapshot percentile distribution is the same every year.

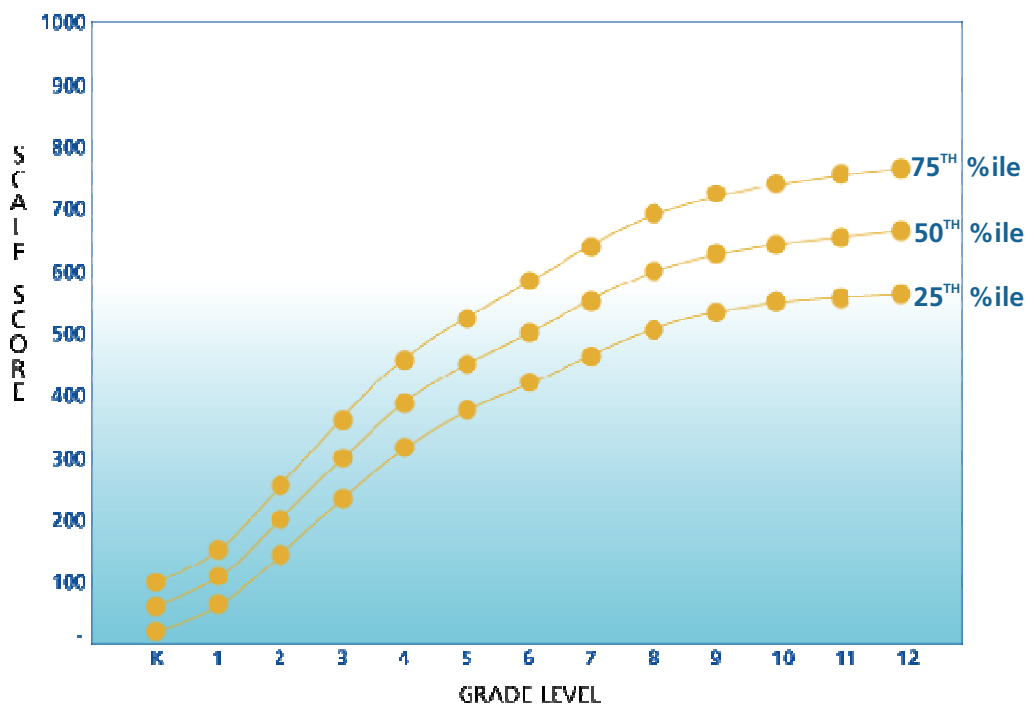


Any or all future students can outperform student in a baseline year.

#### Observation 8. Students diverge.

- As grade levels rise, the gap between low and high achievers widens.
- A high achiever must demonstrate greater than average growth each year to maintain or increase that gap.
- Typical growth for a low achiever allows the gap to increase.

*Why is this important? This is the essence of why some growth models overvalue the growth made by low achievers and undervalue the growth made by high achievers. If a low achiever's growth is only judged within the context of other low achievers, then typical growth will not be enough to close the achievement gap. If a high achiever's growth is only judged within the context of other high achievers, then the success achieved by that student may not be recognized appropriately. This illustrates why the task of closing the gap between low and high achievers is a challenge that gets more difficult each grade level.*



## Current Examples

In North Carolina, an example of what has become typical, the Governor stated that “the number of students performing on grade level increased.” What he referred to was an increase in the percentage of students at the proficiency level or above from one year’s cohort to the next.

### What’s Good:

- Easy to understand
- Grounded in the basic premises of the state’s accountability system
- Based upon the state’s definition of proficiency as being on grade level
- Data/statistics readily available to anyone to verify

### What’s Misleading:

- Very little as long as you know the definition of on grade level and proficiency

In Colorado, “a year’s growth in a year’s time” is defined as achieving a student growth percentile of 50 or higher. The student growth percentile (SGP) is the percentile (rank) of a student’s gain among those other students with the same prior performance.

### What’s Good:

- Defined in detail
- Percentile metric easily understood

### What’s Misleading:

- Based on whether a student outperformed similar students.
- High performer making a large gain equal to similar high performers gets a 50; same as a low performer making a small gain equal to other low performers.
- A 60 (above the average of 50) by a low performer could still mean the student fell farther behind average students.
- A 40 by a high performer could still mean the student gained even farther above average students.
- The complex statistical analysis and formula are not transparent to educators. A complex software program is used to calculate the formula.

What question is the SGP percentile answering?

- Did the student grow as much as other students with the same prior performance level?

Now if you read the first paper on growth, **Growth Model Growing Pains** (available for free download at [www.espsolutionsgroup.com/resources.php](http://www.espsolutionsgroup.com/resources.php)), then you know that this is the value-add question, not the basic growth question. The correction for or leveling of the playing field based upon prior performance means that the “expectation” or criterion for success for each student is ratcheted up or

down based upon whether or not that student is historically high or low performing.

I have no argument with the quantile methodology upon which the SGP is based. However, the final representation of growth should be within the context of all students, not just equally proficient, or non-proficient students. In other words, the question to be asked should have been this one. This is the accountability question that focuses in on whether or not a student is growing at a pace that is generally thought to be average, normal, described in the standards for all students, typical, etc.

- Did the student grow as much as other students?

Simply put, we should not report that a student made a year's growth if that student has fallen farther behind grade level or failed to keep pace with average students. We should also avoid downgrading the success of high achievers by reporting they made a year's growth when in fact they made more than a year's growth merely to maintain their lofty status. What Colorado needs to do is reword its interpretation of the SGPs to be accurate.

## Conclusion

*Educators, Researchers, and Public Information Officers:* If you made it through this paper, then you are fully capable of determining your own definitions. My admonition is that you be accurate in what is reported.

*Parents, Students, and Other Audiences:* Look for definitions of what is reported to you. Insist upon proper use of terms.

Most people are already wary of the tendency to report education data in a positive light. The artifactual growth reported from some growth models exacerbates this perception. The growth models that most frequently represent artifactual growth as true growth are the value-add models. Some models, such as Colorado's student growth percentiles, are value-add models in disguise and can overstate the growth of low performers and understate the growth of high performers in relationship to the proficiency standard established by the state.

These are the definitions that are the most representative of what we as educators, parents, researchers, policy makers, and the general public think we are being told when we hear references to performing on grade level and making a year's growth.

Performing on Grade Level:

- **Standards-Based Perspective:** Grade level is defined as the skills and knowledge established as required for a grade level. The boundary for being on grade level is often referred to as the lowest score that classifies a student as proficient.
- **Normative Perspective:** Grade level is defined as the performance level of the typical student in a grade level. Typical, in a normative sense, is the median or 50<sup>th</sup> percentile student; however, a lower percentile may be used to include all students who might have scored at the 50<sup>th</sup> percentile if retested. In other words, on grade level would include those scoring at or above 50 and all others within some unit of SEM (standard error of measurement) or SD (standard deviation) of 50.

Making a Year's Growth:

- **Standards-Based Perspective:** Maintaining or improving the proficiency level from one year's administration to the next (Maintaining may only apply to students at the proficient level or higher).
- **Normative Perspective:** Making a scale score gain from one year to the next that is equal to or greater than that made by a 50<sup>th</sup> percentile student.







### About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into PK-12 education data systems and psychometrics. Our team is comprised of industry experts who pioneered the concept of “data-driven decision making” and now help optimize the management of our clients’ state and local education agencies.

ESP personnel have advised school districts, all 52 state education agencies, and the U.S. Department of Education on the practice of K-12 school data management. We are regarded as leading experts in understanding the data and technology implications of the **No Child Left Behind Act (NCLB)**, **EDFacts**, and the **Schools Interoperability Framework (SIF)**.

Dozens of education agencies have hired ESP to design and build their student record collection systems, federal reporting systems, student identifier systems, data dictionaries, evaluation/assessment programs, and data management/analysis systems.

To learn how ESP can give your agency *Extraordinary Insight* into your PK-12 education data, email [info@espsg.com](mailto:info@espsg.com).

This document is part of *The Optimal Reference Guide Series*, designed to help education data decision makers analyze, manage, and share data in the 21st Century.

*Performing on Grade Level and Making a Year's Growth – Muddled Definitions and Expectations, Growth Model Series – Part III.*  
Copyright © 2009 by ESP Solutions Group. All rights reserved. No part of this paper shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher.



# ESP Solutions Group

(512) 879-5300

[www.espsolutionsgroup.com](http://www.espsolutionsgroup.com)