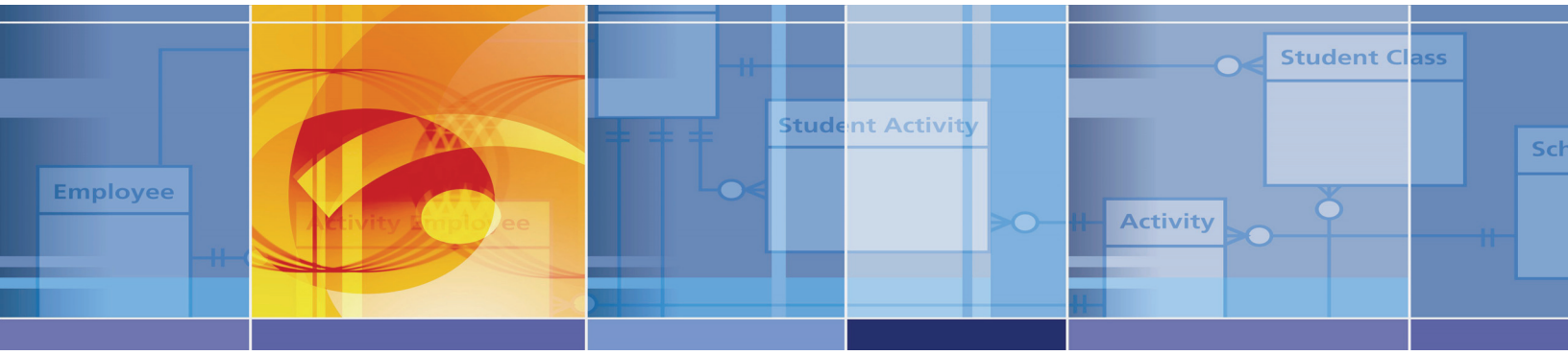


The Optimal Reference Book:

Growth Models – Finding Real Gains

Extraordinary insight™ into today's education information topics

By Glynn D. Ligon

ESP Solutions Group

Table of Contents

Preface

Our series of growth model papers has been a challenge to write. Each paragraph churns up another issue. In the end, we better understand what real academic gain is for a student.

There are really two separate questions that we keep mixing together.

- How proficient is this student?
- How effective is this school?

NCLB requires us to answer the first—and foremost question. In so doing, many educators saw schools being labeled as ineffective—because not enough of their students scored at the proficient level on the annual state assessment.

What We Learned about Growth Models

Stealing the headlines from all the psychometric issues, the popular appeal of growth models has become:

> Accountability systems unfairly label as failing some schools whose students are making academic gains but who are not yet performing at a proficient level.

Setting aside the adverb "unfairly" for now, this statement is true. However, as our series of three papers shows, there are many questions to consider before we understand whether the school is successful.

A Question of Relative Value

As the uses are debated, growth models are currently being over valued, under challenged, and in need of critical evaluation over the next few years. Our series of three growth model papers contributes to the debate and understanding of how we should evaluate the models and, in the long run, select the ones that fit our questions.

How to Use Growth Models

Growth models are a part of the solution in our quest for the answer to how effective our schools are, how effective our educational programs are, and how much our students are learning. In our analysis of growth and value-add models, a natural sequence of questions arises. Following this sequence puts into context the relative importance of each question, and the significance we should place on each.

In the on-going debate about replacing status measures with growth measures, this sequencing of questions makes the point that we should always begin with the status question. The relative unimportance to an individual student of a value-add question and model is also evident.

The Sequence of Academic Performance Questions

For an Individual Student:

1. STATUS — Is the student proficient?

   This is the first and foremost question we want to know. Did the student perform at the proficient level on our measurement?

2. GROWTH — Is the student growing at a pace to be proficient by a target date?

   If the student is not currently proficient, we can have some expectation of future attainment of proficiency if the student's growth is at a pace sufficient to reach proficiency by a target grade level.

3. VALUE ADD — Is the student growing at a pace greater than other similar students?

   If the student is not proficient, and if the student is not growing at a pace to be proficient by a target grade level, then the final avenue for relative positive performance is to out score other similar students.

With all due respect for the complexities that teachers face when diagnosing and prescribing instructional interventions for individual students, here is a simplistic view that illustrates how status, growth, and value-add measures vary in what information they provide. This table shows a diagnosis and prescription for two conditions for each measure.

Diagnosis	Prescription	Measure		
		Status	Growth	Value Add
Student is already proficient and improving faster than other similar students.	Continue current instructional interventions.	Proficient	On Pace	Above
Student is already proficient, on pace to remain proficient, but improving slower than other similar students.	Examine current instructional interventions for appropriateness for this student.	Proficient	On Pace	Below
Student is already proficient, but not on pace to remain proficient, but improving faster than other similar students.	Remedial intervention is needed.	Proficient	Below Pace	Above
Student is already proficient, but not on pace to remain proficient, and improving slower than other similar students.	Remedial intervention is needed; examine current instructional interventions for appropriateness for this student.	Proficient	Below Pace	Below
Student is not proficient, but is on pace to be proficient, and is improving faster than other similar students.	Continue but monitor success of current instructional interventions.	Not Proficient	On Pace	Above
Student is not proficient, but is on pace to be proficient, but is improving slower than other similar students.	Continue but monitor success of current instructional interventions; examine current instructional interventions for appropriateness for this student.	Not Proficient	On Pace	Below
Student is not proficient, and is not on pace to become proficient, but is improving faster than other similar students.	Remedial intervention is needed.	Not Proficient	Below Pace	Above
Student is not proficient, and is not on pace to become proficient, and is improving slower than other similar students.	Remedial intervention is needed; examine current instructional interventions for appropriateness for this student.	Not Proficient	Below Pace	Below

For a School:

1. STATUS — Is the school effective?

   This is the first and foremost question we want to know. Did the school meet the standard for adequate yearly progress?

2. GROWTH — Is the school improving at a pace to be effective by a target date?

   If the school is not currently effective, we can have some expectation of future attainment of effectiveness if the school's growth is at a pace sufficient to reach adequacy by a target year.

3. VALUE ADD — Is the school growing at a pace greater than other similar schools?

   If the school is not effective, and if the school is not growing at a pace to be adequate by a year, then the final avenue for relative positive performance is to out score other similar schools.

Diagnosis	Prescription	Measure		
		Status	Growth	Value Add
School is already adequate, is on pace to stay adequate, and is improving faster than other similar schools.	Continue current instructional programs; be a mentor school for others.	Proficient	On Pace	Above
School is already adequate, is on pace to stay adequate, but is improving slower than other similar schools.	Continue current instructional programs.	Proficient	On Pace	Below
School is already adequate, is not on pace to stay adequate, but is improving faster than other similar schools.	Examine current instructional programs.	Proficient	Below Pace	Above
School is already adequate, is not on pace to stay adequate, and is not improving faster than other similar schools.	Examine current instructional programs with successful mentor schools.	Proficient	Below Pace	Below
School is not adequate, is on pace to become adequate, and is improving faster than other similar schools.	Examine current instructional programs.	Not Proficient	On Pace	Above
School is not adequate, is on pace to become adequate, and is improving slower than other similar schools.	Examine current instructional programs with successful mentor schools.	Not Proficient	On Pace	Below
School is not adequate, is not on pace to become adequate, but is improving faster than other similar schools.	Examine current instructional programs.	Not Proficient	Below Pace	Above
School is not adequate, is not on pace to become adequate, and is improving slower than other similar schools.	Examine current instructional programs with successful mentor schools; review school for closure or restructuring.	Not Proficient	Below Pace	Below

Growth Model Growing Pains, Growth Model Series – Part I

Foreword

Just say growth models in an education agency and the debate begins. These models and their cousins the value-add models embroil us all in one of the greatest politimetric struggles of our time. Overstated? Maybe, maybe not.

Edvance Research, Inc. (ERI) is using a grant from the Michael and Susan Dell Foundation (MSDF) to enable school districts to improve student achievement through the use of leading and lagging indicators. Assessment scores are the quintessential lagging indicator, but growth in achievement is the indicator most highly anticipated by the participants. Can academic growth be measured and reported in such a timely manner that it becomes both a leading and a lagging indicator?

Region 10 Education Service Center (Texas) has built the Empower Data Warehouse with the intent of enabling schools to improve instructional and administrative processes and outcomes through data-driven decision making. They have created a place to gather the data necessary for measuring growth and value-add. They are finding an abundance of models to consider, but await possible legislative action that could mandate an official one. Empower will be ready with the data.

As ESP Solutions Group has worked with our state-level clients, we find the requirement to include data for growth and value-add models to be universal. From Alaska, to Rhode Island, Missouri, Connecticut, Idaho, and many others, we see today's foundation being built at the state level to support expansion into producing reports with whatever growth and value-add models emerge as useful (or mandated). In states like Maine, Delaware, Maryland, North Carolina, Colorado, Texas, California, as well as those cited above, they are building metadata dictionaries and adopting standards to ensure all required data elements are available for state reporting, Federal reporting, and eventually growth and value-add reporting. (Pardon me for not mentioning all the other states we have worked with in these efforts.)

In this three-part series of Optimal Reference Guides (ORGs), growth and value-add models are critiqued, criticized, and praised. Considerably more practical application of these models is needed to know which ones work, how well, and where. Suffice it to say in this foreword that if experts are telling you now that there is a great model you must commit to using, be very cautious.

Now is the time to ensure that your information system contains the longitudinal data that can feed whatever models rise to the top. Limiting your scope at this time to a single approach is risky. Having the capacity to run any and all models for a few years is the way to go.

Introduction

What do people really want to know about growth and value-add models? After reading through the bulk of research and editorials on the subject, and my own practical experience, here are my simple answers. Each is arguable, but then, that's the point of papers like this—to argue, and in the process, discover what we think should be done.

- Do growth models work? Yes.
- Do value-add models work for pay-for-performance plans? Not really.
- Do growth models or value-add models inform instruction? No.
- Do growth models or value-add models function to evaluate program or school effectiveness? Yes.
- Can I, as a typical educator or policy official, understand how most value-add models really work? No.
- Do those statistical formulas like HLM/Mixed Effects tease out any additional meaningful relationships? A few weak ones.
- Do we already really know the low-performing schools and ineffective teachers based upon other data? Yes.
- Will these models help us find some schools that are successful that might be models of best practice? Yes.
- Can we create our own model and compute it, or do we have to buy an expensive one? Create.
- Will schools with low-status ratings turn out to be effective using a growth model? Very few.
- Will a growth model change the AYP status of very many schools that are in need of improvement? No.
- Will many high-status schools be exposed as frauds? No.
- Will this paper inform me or confuse me more? Inform you.
- Do we have all the right, high-quality data needed for value-add models? No.
- What sport should we look to for guidance in discovering the characteristics of a successful school? Baseball.
- Can we calculate the value added by a great librarian? Counselor? No.

Please, don't give up on growth and value-add models completely—yet. Keep in mind that I and others began using growth and value-add models for evaluating school effectiveness in the early 80's. These are not new ideas, but computers make these models more practical today.

When we write these Optimal Reference Guides, crafting a title is one of the more challenging tasks. Many titles were considered for this review of how we should judge academic performance beyond a status measurement. Here are the ones that came in second.

ESP Insight
Please, don't give up on growth and value-add models completely—yet.

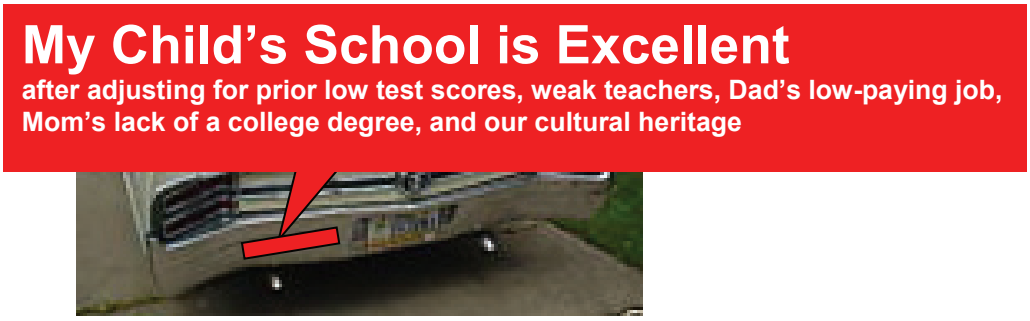
- Transforming Growth Models into Value-Add Models
- Simplifying Growth Models
  —they don't have to be that hard to understand
- Value-Add, Growth, Status?
  —talk like a pro with this easy guide
- Value-Add, Growth, Status?
  —confessions of a user
- Pay-for-Performance
  —why the experts say don't do it
- Why Legislators and School Board Members Don't Seem to Understand Pay-for-Performance
- Abandon Status for Growth, then Abandon Growth for Value-Add, then Abandon Value Add for Legal Reasons
  —the ongoing quest to find the good in every school
- If We Just Measure Growth, No… Make that Value-Add, then Our School Will be Recognized as Successful
- Value *Ad Nauseam*
- Significant or Benign Growth?
- Why Pay-for-Performance Doesn't Work in Public Education
- Is Status the Most Reliable Way to Measure School Effectiveness After All?
- Let the Principals Do It!
- Heisenberg was Wrong When it Comes to Student Assessment
- Perform, Grow, Beat the Odds

- The More We Adjust for Low Test Scores, the More We Excuse Low Performance
- Pay-for-Performance or Pay for Perpetuation?
- No Pass No Pay—if it's good for students…
- What Your Statistician Didn't Tell You about Growth and Value-Add Models

Bumper stickers became popular in the 60's when families got one for every park and attraction they drove to on their summer vacation. Then over the next couple of decades, political statements began to take over the bumper sticker business. Now cars and SUVs cost so much, and their non-metal bumpers are so sticker resistant, that the tradition has faded—but not faded away. Parents have taken up the practice with the new back-windshield, vinyl stickers proclaiming their child's academic and athletic prowess.

Here are some bumper sticker/window sticker slogans that came to mind for growth and value-add models.

- My Low-Performer is Projected to be Proficient
- My Other Student is Proficient without Value-Add
- My Child's School is Excellent
  after adjusting for prior low test scores, weak teachers, Dad's low-paying job, Mom's lack of a college degree, and our cultural heritage



Our exploration of growth and value-add models has been partitioned into three Optimal Reference Guides. The three papers discuss growth and value-add models from very different perspectives. Part I explores the issues related to these models—sometimes with a bit of sarcasm, sometimes with deep respect. This first part is for those readers who like arguing with speeding freight trains. Part II describes the models. This second part is a primer for those wanting to be conversant about models—and which might be appropriate for a particular context. In Part III we review basic concepts about education assessment scores and how we interpret them. Specifically, we examine performing on grade level and making a year's growth.

Do Growth and Value-Add Models Work?

Yes and no. However, you deserve a much more definitive "maybe" than that, so here is a review of the issues:

The National Pastime Model

We are probably looking at the wrong indicators of success.

ESP Insight
Baseball owners discovered that they have been using "official statistics" to judge players, but those statistics don't relate to winning games.

Read *Moneyball* (*Moneyball: The Art of Winning an Unfair Game*. Lewis, Michael. W.W. Norton & Company Inc., 2003. ISBN 0-393-05765-8). Baseball owners discovered that they have been using "official statistics" to judge players, but those statistics don't relate to winning games. For fans, those official statistics are batting averages, home runs, stolen bases, sacrifices, hit-and-runs, and earned run averages. Instead of the official stats, owners should be looking at on-base percentage for hitters (a batter must get on base to score a run) and strike outs/walks/home runs for pitchers (a pitcher shouldn't be judged by runs that score from weak fielding or lucky hits).

The insight from *Moneyball* is that even when the statisticians demonstrated that the new statistics were better, and the Oakland Athletics used them to build a winning team with one of the lowest payrolls in baseball, the other teams stuck to their traditional scouting and drafting practices.

The model from baseball is that if we were to build a value-add model to find successful baseball teams, we could predict the number of games won from the statistics that are most related to winning. Then the teams that are most successfully taking advantage of their resources would win even more games than predicted.

ESP Insight
In education, the talk is more around factors that are out of the control of the students and schools.

One lesson from baseball is that the factors in the prediction model should be those that the players can control. Batters—what's under your control? Pitchers--what's under your control?

In education, the talk is more around factors that are out of the control of the students and schools—prior teachers, prior schools, family income, race/ethnicity, mother's education level, etc. Imagine telling a baseball team owner that you created a value-add model for his team that uses the players' prior coaches, prior teams, salary, race/ethnicity, and mother's batting average.

However, if you tell the owner the model uses prior all-star votes, on-base percentage, times thrown out of a game, and runs scored, the owner might listen. The analogous factors for a student might be prior test scores, attendance, discipline incidents, and grades/credits earned. Now, *those* factors might be predictors of future academic success.

Consider this—in baseball, getting on base, no matter how you do it, is the best predictor of winning games. In the classroom, having a qualified teacher show up to deliver instruction every day is the best predictor of learning. Well, we haven't

quite proven that yet, but evidence is building that teacher attendance is highly significant.

Personally, I believe that in the future we will find more correlates of learning that we either discount or can't measure now. Consider teachers' contact time with students, time on task, time spent grading papers, and time spent giving students feedback. These would be much more interesting factors in a value-add model than prior test scores, race/ethnicity, family income—factors that many people consider merely reasons to excuse poor performance, or to establish lower expectations for some students and schools than for others.

Growth Models are Not Quantum Mechanics

Heisenberg Uncertainty Principle:
In quantum mechanics, the position and momentum of particles do not have precise values, but have a probability distribution. There are no states in which a particle has both a definite position and a definite momentum. The narrower the probability distribution is in position, the wider it is in momentum. Physically, the uncertainty principle requires that when the position of an atom is measured with a photon, the reflected photon will change the momentum of the atom by an uncertain amount inversely proportional to the accuracy of the position measurement. The amount of uncertainty can never be reduced below the limit set by the principle, regardless of the experimental setup.

Heisenberg taught us that we can't measure momentum and position at the same time—at least not without affecting one or the other. Does the same principle apply to measuring student academic status and growth? This is an interesting proposition to explore in our venture to better understand what we are really measuring in our assessment systems.

ESP Insight
We are not really measuring momentum; we are taking sequential status measures and tracking the trend.

Can we measure a student's or a group's academic status and growth at the same time? With all due respect to Heisenberg, yes. The simple reason is that we are not really measuring momentum; we are taking sequential status measures and tracking the trend. Our question is not how fast is the student learning at this time, rather it's how much has the student changed since the last measure.

Measuring student academic progress is not quantum mechanics. We really have no measures of academic pace, rate, or speed. What we have are multiple measures of status from which we infer a rate of learning. Just to be clear, there's nothing wrong with that. A trend based upon multiple status measures is most likely more accurate than a single measure of the pace of learning at one point in time.

I agree with Heisenberg's description of error. We don't increase the error of our growth measure by measuring status because our growth measure is the status measure. Unfortunately, our growth measure compounds the status error. Considering that we compound the error of our measurements across years, it's a wonder that our growth predictions aren't totally off base. However, just by chance, or by the grace of regression to the mean, we are high at times, low at times, right on at times, and on average, are closer than we deserve to be.

Measuring academic growth in education is more like measuring the skills of an athlete. Those skills are best able to be judged within the context of a performance in an event. For example, a runner may be clocked at 13 seconds for 100 meters in middle school, 12 seconds in high school, 11 seconds in college, and 9.69 seconds in the Olympics. We have no comparable measures of performance outside of track meets because the conditions vary.

Why Projections are Tough

Imagine trying to guess which third-grade students will be Olympic champions. Third graders vary widely in their motivation, enjoyment of running, body fat, muscle tone, parental support, economic support, availability of facilities, and on, and on. This is a much better analogy because so many factors go into success as a runner, similar to the number of factors influencing success as a student. A successful runner also relies upon great coaches and trainers.

ESP Insight
So many factors go into success as a runner, similar to the number of factors influencing success as a student.

Definitions

You'll need to understand the differences among the terms used to describe the models. Unfortunately, in this arena, people use different terms and are still making up new ones. This paper will bring some order to the disparate names people have been using. The terms we'll use are defined in Table 1.

Table 1: Definition of Terms

Term	Definition	Example
Measure	A method for describing learning; an outcome measure describes an end point of learning as contrasted with a learning process.	State-Adopted Assessment, Mathematics Section
Score	The performance on a measure expressed as either a discrete category or continuous variable	April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score = 345 and Proficiency Level = Advanced
Status	A score on a measure at a point in time	April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score and Proficiency Level
Growth	The improvement from one status measure to another (Improvement/growth can be a negative value.)	April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score **MINUS** April 18, 2007, Grade 4, State-Adopted Assessment, Mathematics Scale Score

Term	Definition	Example
Value Add	**Unadjusted**: The additional improvement resulting from the positive impact of a teacher, school, or program (without adjustments)	April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score **MINUS** April 18, 2007, Grade 4, State-Adopted Assessment, Mathematics Scale Score **MINUS** Average difference for similar teachers, schools, or programs
	Adjusted: An increase or decrease made to the unadjusted value-add measure to take out the effect of selected factors	Above difference **MINUS** Average difference for similar students, e.g., race/ethnicity, gender, family income, English proficiency, etc.
Projection	The estimate of a score on a measure at a future point in time	Line drawn from April 18, 2007, Grade 4, State-Adopted Assessment, Mathematics Scale Score **AND** April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score **TO** End of Grade 4
Factor	A context characteristic that impacts the score on an outcome measure	Student Factors: Family economic status, race/ethnicity, gender, English proficiency, migratory status, mother's education level, etc. Intervention Factors: Teacher (and teacher characteristics), school, district, program, per-pupil expenditures, opportunity to learn, instructional days in the school year, etc.
Weight	The value multiplied by to get a factor's impact on the outcome measure	Weight for female gender = .003 Weight for limited-English proficiency = -.04 Weight for prior year's score = .211
Model	The formula that relates factors to the outcome measure	

Term	Definition	Example
Pay-for-Performance	A bonus compensation plan that uses outcome measures to determine the dollar amount of the bonus	
Regression, Hierarchical Linear Model, Mixed-Effects Model, etc.	Names for formulas that weight factors for predicting the outcome measure	These are all basically similar algebraic equations with different methods for combining factors to produce a predicted outcome.

The Experts Themselves Doubt Value-Add Models

Value-add is too unreliable to be used for pay-for-performance.

That is my conclusion from reading the collective opinions of the statisticians who write about their favorite models. Even the people who develop the models don't trust them fully. The outlier in this group is William Sanders, who has ventured beyond the assumptions for the data and the limitations of his model. He comes across as believing that the quantity of the data he has processed overcomes the inadequacies of the data and the limitations of the data model itself.

Using value-add models for pay-for-performance will continue to happen, simply because the face validity of acknowledging teachers and schools who overcome the odds to help students is compelling.

The crux of the challenge will come to the front when teachers and schools challenge the precision implied to differentiate between the levels of pay awarded. The models simply are not very precise. The cut-points will have significant error around them.

In the end, I convinced myself—again—that these models do give us another perspective on student learning and school/teacher effectiveness. The reality that frustrates me is that these models compound all that is unreliable about measuring learning. Combine this unreliability with the relatively small differences teased out by most of the statistically complex growth models, and we are back to pondering whether we are being pushed into using these methodologies because they perpetuate the biases of some (influential) people.

ESP Insight
These models do give us another perspective on student learning and school/teacher effectiveness.

Some of the common people-bias statements are:

- Everyone is working so hard that they must be more effective than the tests indicate.
- The tests are unfair to students of this type and background.
- These students really do know this content; they just can't show their knowledge well on a test.

Why Pay-for-Performance Doesn't Work in Public Education

Pay-for-performance works. That's been proven in private industry. Pay-for-performance doesn't work as well in public education because we follow different rules. The much generalized rules for businesses are…

1. Trust your managers.
2. Hold the managers accountable.
3. Give them the money to award.
4. Make the process simple.
5. Accept bias and favoritism.

In contrast, public education follows these rules.

1. Make the system as objective as possible.
2. Make the decisions at a state or district level to minimize the manager's (principal's) influence.
3. Publish the results publicly and compare to be sure everything is equitable.

4. Hire a sophisticated expert to create a statistical model that few understand to rank every student, teacher, and school with large error bands around the individual rankings.
5. Change the system and the rules as people figure out how to game the system.

In private industry, pay-for-performance is either very subjective or strictly by the numbers, e.g., dollars of sales. A sales person may get paid strictly on sales. Other employees are paid on the basis of their bosses' opinion of them. That works because the span of control is limited, and a common process is to give the boss a bonus pool to distribute and step away.

ESP Insight
In a public arena, we demand accountability, equity, and transparency. Spread that demand across thousands of teachers and schools in a district or state, and the pay-for-performance process just breaks down.

In a public arena, we demand accountability, equity, and transparency. Spread that demand across thousands of teachers and schools in a district or state, and the process just breaks down.

The bottom line is that pay-for-performance works best if it's done strictly by the numbers—simple numbers, unadjusted numbers, numbers everyone counts the same way; or subjective numbers awarded by an informed or invested overseer.

If public education were to follow the rules for a successful pay-for-performance plan for teachers, here's how it might look.

Paying Teachers for Performance

1. Annual (or more frequent) measures would be made of student status, growth, and value-add.
2. Principals would be able to access standard reports to see these results.
3. Principals would observe and rate the teachers individually throughout the school year.
4. Principals would publish for teachers in advance how they will rate them for bonuses.
5. Principals would use whatever criteria and additional measures they wish. How the principals weight each measure would be their choice.
6. Principals would follow district or state guidelines for the high, low, and total bonuses allowable.
7. Principals would make their recommendations to their supervisor and defend them.
8. Final amounts would be part of the total compensation of individual teachers and made public.
9. Principals would be held accountable for the overall success of their schools.

Do private industry managers believe their pay-for-performance system is totally fair and honest? Of course not, at least not on a manager-to-manager comparison. Why then do they advocate such a system for public education? To a surprising degree, private industry managers and executives just don't understand the significance of "public" in public education. Legislators generally don't either.

ESP Insight
To a surprising degree, private industry managers and executives just don't understand the significance

The Best Way to Determine Dollars for Performance

Why not, if we trust our scale to differentiate at all, mitigate our error by giving precise amounts instead of compounding our error by chopping teachers into large groups that reward all teachers within the group identically and make huge pay differences between those at the margins?

ESP Insight
If we pay precisely for the scale score, then knowing a teacher's compensation will also reveal the teacher's rating. Well, why not?

Since the major legal and acceptance hurdle for a pay-for-performance plan is that tricky cut-point between those that get money and those that get nothing at all, here's my strawman solution. Instead of trying to make precise slices between the groups of employees that deserve $3,000 and those that get $0, let's make the scale continuous. With any pay plan, we must place the teachers along a scale. So instead of dividing that scale into categories of performance, we merely pay a certain number of dollars for each point along the scale. For example, an amount of $10,000 is established as the reward for the very best teacher—rated 1.00 on the scale. A teacher with a rating of .234 would get $10,000 times that rating = $2,340. A teacher at .000 or below gets nothing. With this approach, the two teachers who might otherwise have occupied the scale scores on either side of the cut-point with one getting $3,000 and the next getting $0—an arbitrary and indefensible differentiation—would get virtually the same dollars.

Instead of paying all teachers in a wide range the same dollars, the two teachers who occupy the highest scale point and the lowest scale point in the same range would get significantly different amounts.

Of course, here comes the old problem with being in a public agency. If we pay precisely for the scale score, then knowing a teacher's compensation will also reveal precisely the teacher's rating. Well, why not? That's what we must buy into when we use pay-for-performance in a public school setting for a large group of professionals with the same job classification.

Let's look at some sample pay-for-performance tables and react to a crucial issue with pay-for-performance. Do we really think the measures are precise enough to make these decisions?

Assume that whatever model is used, the performance ratings are expressed as a value from 1.000 being the highest possible to -1.000 being the lowest. There is no need to assume a normal distribution of the ratings.

Table 2 presents a typical pay-for-performance table with three categories associated with different compensation amounts.

Table 3 presents a continuous payment scale with each point on the ratings scale associated with a different compensation amount.

A major issue for education agencies defending their pay-for-performance plans is exposed by looking at the dividing line between those getting some dollars and those getting none. In the categorical plan (Table 2), a teacher rated as .001 gets $3,000. A teacher rated .000 gets $0. However, in the continuous plan (Table 3), these two teachers get $10 and $0 respectively. A teacher might fight for $3,000 but might just be quiet over $10. The reliability of the difference between these two teachers' ratings is hardly defensible.

At the upper end of performance, the same holds true. In the categorical plan, a teacher rated as .500 gets $5,000. A teacher rated .499 gets $3,000. However, in the continuous plan, these two teachers get $5,000 and $4,990 respectively. The continuous plan stays out of court again.

ESP Insight

A major issue for education agencies defending their pay-for-performance plans is exposed by looking at the dividing line between those getting some dollars and those getting none.

A major issue with categories (even with student performance categories) is that the highest scorer and the lowest scorer in the same category get the same recognition—even though the difference between the two is much larger than that between many of the individuals on each side of a categories boundary. For example, in the categorical plan, a teacher rated as 1.000 gets $5,000. A teacher rated .500 gets $5,000. However, in the continuous plan, these two teachers get $10,000 and $5,000 respectively.

Table 2: Categorical Pay-for-Performance Plan

Rating	Bonus
1.000	$5,000
.900	$5,000
.800	$5,000
.700	$5,000
.600	$5,000
.500	$5,000
.499	$3,000
.400	$3,000
.300	$3,000
.200	$3,000
.100	$3,000
.001	$3,000
.000	$0
-.100	$0
-.200	$0

Table 3: Continuous Pay-for-Performance Plan

Rating	Bonus
1.000	$10,000
.900	$9,000
.800	$8,000
.700	$7,000
.600	$6,000
.500	$5,000
.499	$4,990
.400	$4,000
.300	$3,000
.200	$2,000
.100	$1,000
.001	$10
.000	$0
-.100	$0
-.200	$0

Other Things That Need to Change

Public education depends upon its academic assessments (statewide tests) as the most affordable, reliable, valid, objective, and comparable measures. A major shortcoming in the models being used and proposed is that they do not account for the number of instructional days (or hours) between administrations of measures. They allocate all or an arbitrary proportion of growth to a teacher or school irrespective of the actual time the student spent in the school.

ESP Insight
A major shortcoming in the models being used and proposed is that they do not account for the number of instructional days (or hours) between administrations of measures.

For a start, these changes are needed.

1. Expand the window for assessments, but begin calculating growth by the number of instructional days between administrations of the measures.
2. Measure students when they enroll in a school/class or as near the first day of a school year as possible.
3. Measure students when they withdraw from a school or as near the end of a school year as possible.
4. Revise information systems to document instructional time, days, and teachers responsible for instruction—by date.
5. Revise the assessments to be "focused accountability measures" rather than "hybrid diagnostic/accountability tests." See ***Why Eva Baker Doesn't Seem to Understand Accountability—The Politimetrics of Accountability,*** ESP Optimal Reference Guide, 2008 (available for download at www.espsolutionsgroup.com/resources.php).

The Socio-Economic Mandate

Before you read this, keep in mind that this is intended as a thought-provoking idea, not an opinion. To hear some people discuss assessments and growth models, it seems that their criterion for validity must be:

ESP Insight
The fact is, everything doesn't matter.

- Face validity for a test is that the privileged students outscore the underprivileged.
- Face validity for a growth or value-add measure is that some high-performing and some low-performing schools move to the middle of the ranking.

Reality may differ.

Statisticians Make Growth Models Too Difficult

That's especially true for value-add models. No, it's not for job security. That's just how they think. They jump to the most complex model because it considers "everything," and education audiences typically want everything explained. The fact is, everything doesn't matter. What might seem to matter may not make that much difference in a statistical model. For example, a value-add model may include race/ethnicity as a factor, but the weight or statistical importance of that factor may be insignificant if the model also includes a reliable measure of socio-economic status. In the world of politimetrics, keeping race/ethnicity in as a factor may be necessary, even though it adds little or nothing to our understanding of school effectiveness—in a mathematical model.

In the end, we want a model that we can understand—one we can defend. This is because when we understand a model, we can do something with the data--use them to improve learning. We appreciate what makes a difference, and what doesn't.

ESP Insight
In the end, we want a model that we can understand—one we can defend.

Even more, with a value-add model that is to be used for pay-for-performance, the model must be explainable and defendable in a courtroom.

Is NCLB Getting Growth Right?

The most recent guidance from USED allows growth sufficient to place the student on pace for proficiency by the highest grade level of the current school to be counted in the proficiency category is certainly consistent with the intent of the No Child Left Behind Act (NCLB). This acknowledges that 100% proficiency by 2014 can be claimed as long as every student is at least on pace for proficiency.

Actual counts by states using this new growth alternative are not finding many schools moving from unacceptable to acceptable by virtue of growth.

Growth Models Won't Save Many Low-Performing Schools

Too bad, but most really low-performing schools look miserable no matter how we analyze their data. In fact, most high-performing schools continue to look good even when we use fancy models to take away their perceived advantages of having high-income, non-minority, native English speakers, and the best teachers. What we are really looking for are those outliers, the schools that teach their students more than similar schools would regardless of what model jalopy, SUV, or bus drops them off at the curb.

ESP Insight
The ultimate irony of a growth model is that if it is effective in informing instruction and improving learning, then it will over time become useless as a predictor of outcomes.

We Want Our Models to Fail—Eventually

The ultimate irony of a growth model is that if it is effective in informing instruction and improving learning, then it will over time become useless as a predictor of outcomes. Think about the dynamics of intervening in instruction based upon the outputs of a growth model. If we improve the performance of low performers by changing processes or resources because a growth model has targeted them for improvement, then those factors will be taken away as things that differentiate poor performers from high performers. The growth model will no longer be predictive. That's great, by the way.

This is not a catch 22, because as one predictor begins to fail us, others are likely to emerge. If not, then we've fine-tuned our instructional delivery processes to their maximum effectiveness.

For example, NCLB hopes that in 2014 every student will be proficient--some after no more than one full academic year in a school. If this were to come about, our models would need to focus more on backsliders than on students with long-term trends of low performance. Those bubble students who barely made it into the proficient category present a risk to their schools that they might slip just a bit and no longer be proficient the next time they are assessed.

Questions for Your Growth Model Expert

As someone who at times assumes the role of expert, I appreciate the trust clients invest in an expert. However, like teachers and principals, not all experts are equally effective. So, here's my list of questions I would insist on being answered by someone proposing a growth model to my education agency.

1. What assumptions does your model make about our students, our schools, our data, and our assessments? How closely are those assumptions met? For example, the model may assume that students are randomly assigned to a school each year from the entire population of students in a state. Yes, some models use sampling statistics as the basis for their analysis. I haven't found a school yet that meets that assumption. The model may also assume that the cut point for proficiency at each grade level is of equal difficulty. By the way, what does that mean?

2. In terms of scale scores on my assessment, how much difference does each predictor variable really make? Ask for a translation of the influence of each factor in terms of scale score points—or fractions thereof. Don't be awed by beta weights or significance levels. With large groups, very insignificant values can be statistically significant.

3. Will you bet your consulting fee on your predictions being accurate? Will 95% of the students you say will be proficient be proficient in the future? This is a sucker bet.

4. In the end (the school's highest grade level, e.g., 4, 5, 6, 8, 12), will the students your model declares as "successful" today be proficient?

5. Can you explain your model to me without using any of these words? Hierarchical, regression, variance, standard error, black box, mixed effects

6. What do you do when data are missing? What do you do with new, mobile students? What do you do with grade levels not tested?

7. Are the predictors you use ones that can be influenced by the schools and people being held accountable?

ESP Insight

Are the predictors you use ones that can be influenced by the schools and people being held accountable?

8. Do the periodicities of the outcomes align with the periodicities of the predictors? Do the testing dates between which growth is measured line up with the same time period when the teacher or school was responsible for the student and able to influence learning? Are the skills being measured the same ones that were to be taught during the time measured?

9. Are your predictions within the standard error of measurement (SEM) of the assessments? If the SEM of the assessment is larger than the difference between a student's score and the cutpoint for proficiency, then the conclusion about the student could change simply by retesting.

10. When can your model be run to get new results? Are you stuck waiting a full year to see the next results?

11. How do predicted results look compared to a prior cohort's results? If the model is predicting that 88% of the fourth graders will be proficient by grade 8, are 88% of the current eighth graders proficient? If the model were to be backdated for the current eighth graders, would the ones predicted to be proficient be the one who became proficient?

Thinking Inside the Box

We describe the growth models that we don't understand as black boxes. There are other analogies that are helpful.

Black Boxes

We really know what happens inside the black box of even Sanders's closely guarded model. The secrets are the actual calculated values and weights for the components of the model. The fact is, those are the numbers that really determine if the model is effective or even being calculated correctly. MORE DATA = MORE STABILITY

Gray Boxes

The box is gray not because the values and weights are secret, but because if the typical educator saw inside the box, everything would still be a mystery. Gray boxes are transparent, but not easily understandable.

Transparent Boxes

The box is transparent when the typical user of the model can look inside and follow what's happening. Maybe that user can't perform the calculations, but the steps and the impact of those steps are understandable.

Caves

The analogy to a box is fine, but instead of discrete boxes, our growth models fall along a more continuous line from simple to complex. A better analogy might be a cave. The simplest models reside near the entrance to the cave, and a user can peer in and see everything fairly brightly. As models get more complex, they move farther into the cave and are less illuminated. In the back of the cave, the most complex models may be totally dark for a user.

I like the cave analogy because as the data emerge out of the cave, we can begin to interpret them. The darkest models from the back of the cave require the longest journey to reach a level that is illuminated. At the entrance to the cave, where the user stands awaiting the results, everything ends up abstracted to the same level for interpretation, regardless of the fancy calculations conducted in the dark recesses of the cave.

Limitations on Growth and Value-Add Models

Take a deep breath before reading this section. You must be passionately committed to growth and value-add models to survive the litany of shortcomings and problems they face. For me, these issues do not out weigh the benefits of having growth and value-add information in our D3M processes.

ESP Insight

You must be passionately committed to growth and value-add models to survive the litany of shortcomings and problems they face.

1. Measuring growth requires that we have multiple years of assessments, preferably consecutive years.
2. Few education agencies have acceptable assessments before the end of grade 3.
3. Few education agencies have acceptable assessments after grade 10.
4. Student Mobility reduces the number of students with multiple assessment scores.
5. Student mobility reduces the number of students with consecutive years of enrollment within the same school.
6. Teachers and staff change within schools.
7. School and district demographics change.
8. Assessments change from year to year.
9. The skills measured vary across grade levels.
10. The alignment of skills with the assessments varies across grade levels.
11. The test might be scheduled before the skills are scheduled to be taught.
12. Opportunity to learn the skills tested may not be afforded.
13. Assessments have caps that limit the measurement of growth for high achievers.
14. Assessments have floors that limit the measurement of true starting performance for low achievers.
15. Regression toward the mean lowers growth.
16. Regression to the mean raises growth.

17. Agencies typically use internal consistency measures of reliability rather than test/retest measures that produce larger measurement error estimates. (Thanks, James Popham, for teaching and reteaching me this.)

18. SEM limits the accuracy of a growth model.

19. Value-add models may include factors that have little or no correlation with the assessment.

ESP Insight
Let's demand that the results be reported by subject area and grade level, rather than aggregated across an entire school.

20. The reliability of the measurement of each factor lowers its predictive power.

21. Vertical scaling of the assessment may be lacking or only estimated.

22. Assumptions of the analysis technique may not be met by the data or the assessment.

23. Missing data changes the evaluation question being answered.

24. The proficiency cut points at each grade level may vary in their difficulty.

25. Attribution of growth to a teacher, school, or program may have competing factors.

26. Attribution of learning to staff other than teachers is difficult.

27. In-school vs. out-of-school factors compete for causal relationships.

Be wary of experts who are dismissive of the limitations of their models.

The Question Parents Should be Asking

Which school is the most effective with students like my child?

Although a school might be extremely effective with high-income, majority students who are college bound, it may have little to offer students who differ from this demographic. On the other hand, a school that is the very best at teaching limited-English proficient students may be lacking in effective strategies for native-English speakers with mathematics needs.

To pile on with demands for any growth or value-add model, let's demand that the results be reported by subject area and grade level, rather than aggregated across an entire school.

How Many Growth and Value-Add Models are There?

Fewer than we typically think.

Part II of this paper will detail those. I'm open to debate on the count, but it may be as few as three.

The conclusion in Part II is that value-add models are simply a case within each of the growth model types. Interestingly, some people lump all value-add models into one type. They don't discuss the process of turning a simple growth model into a value-add model by merely including a comparison group or expected score.

Some of the confusion in this arena comes from the creative names people have given to their models. Even more arises from the idea that there is a different model if the calculation uses a different metric, for example vertical scale score vs. percentile rank.

Growth vs. Value-Add

1. The only practical difference between a "Growth" model and a "Value-Add" model is that to be value-add, the model must "control for" the influence of selected factors (e.g., demographics, prior performance, etc.) or the impact of an intervention (e.g., school, program, teacher, etc.) on student performance.
2. Controlling for these factors allows the interpreter of the results to say, "The growth beyond what was controlled for represents the value added by the school, for example."
3. When value is added by a school, the school is said to be effective—regardless of the status of students' performance.
4. Value-add models are used for several purposes beyond description of student performance.
    a. Pay-for-performance: When teachers or whole schools are rewarded for growth above and beyond what would have been predicted by the factors in the model.
    b. Evaluation: When a program is deemed effective because its students outperformed other similar students.
5. For this review, growth models and value-add models are analyzed together because growth models are merely a simple, unadjusted case of the same models that are labeled value-add.

ESP Insight

The only practical difference between a "Growth" model and a "Value-Add" model is that to be value-add, the model must "control for" the influence of selected factors.

ESP Insight
Value-add models come with a curse.

The Value-Add Conundrum

Value-add models come with a curse. "Well, that's the best we can expect of those students considering their past performance and demographics." Some accountability systems have purposely chosen not to use a value-add approach because one might be equated with an excuse for low performance by minority or low-income groups. NCLB is the best example. The goal in 2014 is 100% proficiency. Even with growth models approved by USED, the growth must be sufficient for the student to reach proficiency by the end of a school's grade span or 2014.

How can an accountability system avoid the criticism of low expectations while implementing a value-add model? A combination of metrics must be reported.

1. Status. Yes, report the actual status to be clear whether or not the students overall are performing at an acceptable current level.
2. Growth.
3. Value-Add.

What's the Criterion for Gain?

This is the issue that challenges all growth and value-add models. We can calculate the numbers. Interpreting them is more difficult—especially with the value-add conundrum looming.

Some possibilities are:

- The student moved up.
- The student moved up more than others.
- The student is on pace to be proficient by___.
- The student moved up an established amount.

Avoiding Leaving a Child Behind

Averages do mask the low performance of some students with the high performance of others. They also, mask the high performance of some students with the low performance of others. Growth and value-add models that produce averages are prone to following this path of obfuscation.

The remedy is to use an individual student approach for the model and report the number and percent of students by performance categories rather than an average.

The Bottom Line on Pay-for-Performance

A principal's judgment of teacher performance is biased. The results of a value-add model for a teacher's effectiveness is unreliable. A legislature that substitutes a principal's professional judgment for the error in a statistical model is expressing an unfortunate distrust of our principals, and an uninformed confidence in psychometrics.

I say hold the principals accountable at the school level. Give them the leeway to use pay-for-performance to reward their best staff members—with the full knowledge that their own accountability rests on the outcome of the status, growth, value add process.

One fatal flaw in assessment-based pay-for-performance in public education is the limited number of student cases available for determining an individual teacher's value. Let's review some of the well-known shortcomings.

1. Elementary teachers of record may or may not deliver instruction in the areas measured by the test.
    a. Only 30 students a year is a small number for reliability.
    b. 90 students over 3 years is better, if the test and academic standards remain constant.
2. Not all secondary teachers have responsibility for courses measured by the tests.
3. Teachers move around in their schools and responsibilities.
4. Schools have about as many other staff and employees as they have teachers who have direct responsibility for instruction.
5. Teachers may be hamstrung by the textbooks, curriculum, supplemental materials, and assigned lesson plans they must use.

Awarding the Proper Value for Value-Add

I worry that to the casual observer attaining a positive level on a value-add measure will be interpreted as better than it really is. The extreme example follows.

> The school exceeds its expected level of performance on the value-add measure by outscoring other schools with the same characteristics. Enter celebration and recognition. However, the school is below the goal for student performance, and the school's growth trend is too slow to reach the goal by a reasonable target date.
>
> Proper interpretation is that the students in this school are not only performing low today, but they will end their education performing below what has been established as the standard to meet. However, because they outperformed other unsuccessful students, we consider them successful.

I also worry that to the casual observer (aka legislator, senator, congresswoman, or congressman) attaining a negative level on a value-add measure will be interpreted as worse than it really is. The extreme example follows.

> The school misses its expected level of performance on the value-add measure by not outscoring other schools with the same characteristics. Enter dismay and withholding rewards. However, the school is above the goal for student performance, or the school's growth trend is fast enough to reach the goal by a reasonable target date.
>
> Proper interpretation is that the students in this school are not only performing well today, but they will end their education performing at or above what has been established as the standard to meet. However, because they did not outperform other successful students, we consider them unsuccessful.

Emotions have entered this arena. Some people resent the "free ride" they see high-income, high-performing schools getting because their students enter achieving well and leave achieving well regardless of whether or not the school is effectively teaching them. On the other hand, some people resent the "losing cause" they see low-performing schools having because their students enter achieving poorly and leave achieving poorly regardless of whether or not the school is effectively teaching them.

With the reality of the ceiling and floor effects of our assessments, the loss of scores for mobile students, and all the other challenges we have implementing longitudinal measurements, we must acknowledge that we cannot precisely determine reality. Our assessments may not be sensitive enough to growth for high-achieving students. Low-performing students, especially those of initial limited-English ability may make progress imperceptible to our assessments.

ESP Insight
Our assessment may not be sensitive enough to growth.

Not to belabor these points, let's move on to a solution.

Tiered Performance Indicator

The foundations of this indicator are:

1. The prime objective is for each student to meet or exceed the performance goals before graduation.
2. Three conditions can exist for a student:
    a. Currently meting this objective
    b. On track to meet this objective
    c. Not on track to meet this objective
3. A school's rating is based upon how many students are in categories a plus b.
4. Value-add is calculated for failing schools as information to assist in determining the proper school improvement interventions.
5. Value-add is calculated for successful schools as information for identifying best practice campuses.

As you can see, emphasis is still on status performance. Growth is acknowledged if the trend is sufficient to achieve the goal status by the target year. Value-add is only used to identify best practices or to inform decisions about how to improve failing schools.

ESP Insight

Value-add is only used to identify best practices or to inform decisions about how to improve failing schools.

This methodology will be expanded in Part III.

Conclusions

Maybe these are more opinions than conclusions.

- Growth and especially value-add models are over-rated.
- However, these models add to and help balance our perspective on student learning, and teacher/school effectiveness.
- Some growth and value-add models are more complex than their results justify.
- As a decision-maker, I want status, growth, and value-add information.
- However, as a taxpayer, I want limited use of value-add in a pay-for-performance plan.

Comparison of Growth and Value-Add Models, Growth Model Series – Part II

Foreword

There are only a few really different approaches to growth models, but many different formulas for calculating them. If we understand which question each model answers, then making a selection among them will be easier.

This paper will examine, at a high level, the characteristics of different growth and value-add models. If you have not at least skimmed Part I then you should go back and do so. This paper is called Part II for a reason. This paper starts beyond the basics of growth measurement and assumes the caveats, limitations, and admonitions about them are understood.

We should define the term "model" as it is used in this paper. Model here means a high-level category of statistical techniques that answer a specific question for specific purposes. This paper does not give formulas or recommend specific statistical techniques (too often called models by their authors or proponents). The growth and value-add solutions being marketed can be used for one or more of the models described in this paper.

Background

Value-add models are simply a case within each of the growth model types. Interestingly, some people lump all value-add models into one type. They don't discuss the process of turning a simple growth model into a value-add model by merely including a comparison group or expected score.

Some of the confusion in this arena comes from the creative names people have given to their models. Even more arises from the idea that there is a different model if the calculation uses a different metric, for example vertical scale score vs. percentile rank.

Growth vs. Value-Add

1. The only practical difference between a "Growth" model and a "Value-Add" model is that to be value-add, the model must "control for" the influence of selected factors (e.g., demographics, prior performance, etc.) or the impact of an intervention (e.g., school, program, teacher, etc.) on student performance.

2. Controlling for these factors allows the interpreter of the results to say, "The growth beyond what was controlled for represents the value added by the school, for example."

3. When value is added by a school, the school is said to be effective—regardless of the status of students' performance.

4. Value-add models are used for several purposes beyond description of student performance.

    a. *Pay-for-performance:* When teachers or whole schools are rewarded for growth above and beyond what would have been predicted by the factors in the model.

    b. *Evaluation:* When a program is deemed effective because its students outperformed other similar students.

5. For this review, growth models and value-add models are analyzed together because growth models are merely a simple, unadjusted case of the same models that are labeled value-add.

The Model Comparison Chart

Describing the models and their characteristics in text became convoluted, so a chart was created to display all the information together. See the Model Comparison Chart at the end of this paper.

1. **Our Questions**

   The chart begins by stating the question asked about student performance, school effectiveness, or teacher effectiveness. Each question is composed for growth and for value-add.

2. **Related Evaluation Questions**

   Each question is restated to be measurable for groups or individuals.

3. **Model Names**

   The model names come from a review of the literature on growth and value-add models. From all the references, several were selected because they describe similar models to the one being described in the chart. The researcher or practitioner to whom the model name is attributed along with the model name they use are listed to help consolidate all the various terms being used for the same model.

4. **Use**

   Some models are better suited for different purposes. The preferred uses are shown for each.

5. **Pre-Measure**

   Acceptable metrics usable as pre-measures are shown as assessment scores by year and grade level. For group measures, the sequence of the cohort is indicated. For individual measures, the student is shown.

6. **Post-Measure**

   Acceptable metrics usable as post-measures are shown as assessment scores by year and grade level. For group measures, the sequence of the cohort is indicated. For individual measures, the student is shown.

7. **Calculation**

   The appropriate statistical calculation(s) is indicated using the pre- and post-measures.

Statistical significance tests are not included because they must be matched to the assumptions of the data used within each model. For example, interval data may use means and parametric tests; whereas, categorical data require non-parametric analyses, possibly employing the standard error of measurement.

For individual measures, the standard error of measurement is the metric of choice for determining an error band or confidence interval.

"School," as used in this comparison, may be a program, intervention, resource, or other factor thought to impact student learning outcomes.

Factors (covariables or predictors) for value-add formulas are listed on the chart for students, teachers, and schools.

The Real Growth and Value-Add Question

The questions require one additional modification. The stated questions apply only to those students with assessment scores. Unlike a status measure that under NCLB represents at least 95% of the eligible students, a longitudinal measure includes only those students with multiple measurements.

Each question could (should) be preceded by "For those students with whom the school had the opportunity to instruct for all the years included in this analysis…" Obviously, no growth model has data to measure the impact of the school on students that come and go within the timeframe of the model.

Is that really true? Keep in mind that some instances of these models (e.g., SAS) impute missing values. This means that they move forward with gaps in data for some students by plugging in an estimated value. Not a dreadful strategy, but users of that model must understand that for those students, the predicted or estimated performance has more error in it than for other students. (i.e., a combination of measurement error and sampling error).

With the emergence of universal statewide assessment programs, tracking student learning has improved. Thank you, NCLB. Today, a mobile student may bring along a pre-measure from another district in the same state—maybe even multiple years of pre-measures. That bonus adds to the number of students that can be included in an analysis.

How Many Years of Data Are Required for Each Model?

One might think that all models require more than one year of data for every student. In fact, the quasi-longitudinal model is a comparison of two status measures for different cohorts of students. Only one year of data is needed for each student because there is no expectation that the same students are in each year's cohort. This is the only model that includes all students tested in a year. (See **Models that did not Make the Chart** for possible exceptions to this statement.)

The *longitudinal model* requires at least two years of data for each student included.

The *trend comparison model* requires three or more years of data for each students.

The *trend-to-target model* can work with only one year of data, but the more years in the calculation, the better the prediction is assumed to be. Having more years of data in the calculation is a statistical plus for the model. However, changes in the assessments or academic standards being taught over the years are more likely with more years included.

Models that did not Make the Chart

Maybe there is another value-add model. Let's call it the quasi-value-add-model. Some evaluators and researchers have used a shortcut to growth and value-add measurement that requires only a single assessment score. Using a regression approach, the single score can be predicted, post-hoc, from demographic, context, and assessment data other than from the outcome measure. This is really a prediction of status, but it certainly meets the basic criteria for a value-add model.

In fact, several states have reported their annual status averages for schools in comparison groupings of similar schools based upon wealth, percent minority, percent economically disadvantage, etc. This quasi-value-add-status model is also represented by Just for the Kids, which reports a school's performance in comparison to the top performing schools with similar demographics. Any educator who runs a query on an assessment database to compare assessment performance across schools or districts of similar characteristics is running a quasi-value-add-status model.

These are certainly not growth models, but growth above or below some comparison is implied in the result.

I can support the use of these models for accountability, research, and evaluation. The reason—the extra information a decision maker receives from a growth or value-add model far out weights those limitations.

We must however, be ever vigilant that those decision makers understand the limitations—especially the real questions being answered. Even as a decision maker might be misled by only knowing the status of a school, that decision maker might also over interpret a positive growth result—especially a positive value-add result.

Over confidence in a value-add result is easy to understand. After all, value-add formulas are complex and they take into account so many factors that make comparison of status measures unfair. True, but they also impose different expectations for learning on low- and high-achieving students.

The Metric Metaphor Misconception—Metaphorically Vertical Scale

Many statisticians and psychometricians insist that an assessment must have a true vertical scale in order to properly measure growth. I happen to side with the others who are comfortable with the idea that the statistical models do not require either the predictor variables or the predicted variables to be on the same scale. After all, the non-assessment factors used in value-add models are all on different scales (e.g., family income, race/ethnicity, gender, age, etc.).

I have developed a liking for several metaphorically vertical metrics that should perform even better than others. The premier option is the true vertical scale. When there is none, an estimated vertical scale can work. These include:

- Standardized score within each grade level
- Percentile within each grade level
- Normal Curve Equivalent within each grade level

Selecting a Model

This chart will help in determining:

a. What your question is
b. What use you are pursuing
c. Whether or not you have the data to calculate the gains or value-add

Next, you call a statistician. Fortunately, with the issues detailed in Part I and the characteristics in Part II's chart, you have a reasonable opportunity to understand whether or not the statistical approach recommended by the statistician will meet your real needs and wishes.

Now, you are ready to go back, re-read, and even argue with Part I.

Model Comparison Chart

ESP Insight

There are only a few really different approaches to growth models, but many different formulas for calculating them. If we understand which question each model answers, then making a selection among them will be easier.

Our Questions	Restated Evaluation Questions		Model Names (Attribution) [Number of Measurements]	Use
	Groups	Individuals		
Growth: Did this year's students score higher than last year's students? **Value Add:** Is this school more effective than it was last year? Is this teacher more effective than he/she was last year?	**Growth:** How does this year's cohort compare to the prior year's cohort in the same grade level? **Value Add:** How does this year's cohort compare to the prior year's cohort in the same grade level after adjusting for selected factors or in comparison to a similar group?	Not Applicable	Quasi-Longitudinal (Ligon) Improvement (Hull) Successive Groups (Carlson) Growth Relative to Others (Gong) [One measurement per student; only model that includes all students tested]	**Growth:** School Improvement AYP Safe Harbor Accreditation **Value Add:** Pay for Performance
Growth: Did this year's students score higher than they did last year? **Value Add:** How effective was this school? How effective was this teacher?	**Growth:** How much did the cohort's performance change? **Value Add:** How much did the cohort change after adjusting for selected factors or in comparison to a similar group?	**Growth:** How much did the student's performance change? **Value Add:** How much did the student change after adjusting for selected factors or in comparison to similar students?	Longitudinal (Ligon) Simple Growth (Hull) Longitudinal (Carlson) Growth Relative to Others (Gong) Eductional Value-Added Assessment System (Sanders) Dallas Value-Added Accountability System (Webster) [Two measurements per student]	**Growth:** School Improvement Evaluation Accreditation **Value Add:** Pay for Performance Evaluation Research
Growth: Did the students learn at a faster pace this year than they did in the past? **Value Add:** Was this school effective in increasing the students' rate of growth? Was this teacher effective in increasing the students' rate of growth?	**Growth:** How did the cohort's trend in growth change? **Value Add:** How did the cohort's trend in growth change after adjusting for selected factors or incomparison to a similar group?	**Growth:** How did the student's trend in growth change? **Value Add:** How did the student's trend in growth change after adjusting for selected factors or in comparison to similar students?	Trend (Ligon) Change in Rate (Carlson) [Three or more measurements per student]	**Growth:** School Improvement **Value Add:** Pay for Performance Evaluation Research
Growth: Will these students be proficient by the time they leave this school or by 2014? **Value Add:** Is this school effective in increasing or keeping all students' learning on pace to be proficient by the target date? Is this teacher effective in increasing or keeping all students' learning on pace to be proficient by the target date?	**Growth:** Is the cohort on track to reach proficiency by the target date? **Value Add:** Is the cohort on track to reach proficiency by the target date after adjusting for selected factors or incomparison to a similar group?	**Growth:** Is the student on track to reach proficiency by the target date? **Value Add:** Is the student on track to reach proficiency by the target date or in comparison to similar students?	Target Date (Ligon) Growth to Proficiency (Hull) Growth Relative to a Standard (Gong) Eductional Value-Added Assessment System (Sanders) Dallas Value-Added Accountability System (Webster) Rate of Expected Academic Growth (Doran) [One or more measurements per student]	**Growth:** School Improvement AYP Growth Accreditation **Value Add:** Pay for Performance Evaluation Research

NOTES:

1. "School" can also be program, intervention, resource(s), or other factor thought to impact student learning.
2. Sources for model names are:
   a. Ligon, Glynn, Typical research references
   b. Hull, Jim, Center for Public Education, National School Boards Association
   c. Carlson, Dale, Assessment Consultant
   d. Gong, Brian, Center for Assessment
   e. Sanders, William, SAS
   f. Doran, Harold
3. Acceptable metrics may be referred to by different commercial or localized names that are equivalent to the textbook names used here.
4. Questions imply that they are for "students who have been enrolled for the time period being measured."
5. SEM (standard error of measurement) is the estimated standard deviation of the error in that method.
6. NCE is the old Chapter I normal curve equivalent.
7. Parametric refers to statistical techniques used on equal interval data—means, standard deviations, analysis of variance, etc.
8. Nonparametric statistics are medians, etc.

FACTORS (COVARIABLES, PREDICTORS) FOR VALUE-ADD FORMULAS:

Students: Prior scores for the same subject area, combined prior scores for multiple subject areas, teacher(s) of record, teacher(s) delivering instruction in a subject area, individual demographics (e.g., age, gender, race/ethnicity, socioeconomic status/economic disadvantaged status, census track, parent education level, etc.), program participation (e.g., Title I, ESL, Special Education, etc.), learning abilities/disabilities (e.g., handicapping conditions, 504, gifted/talented status, etc.), prior academic success (e.g., grades, promotion/retention, honors, etc.), discipline incidents, attendance, school of enrollment, district of enrollment, state of enrollment, school feeder pattern, instructional intervention/pedagogy, benchmark assessments, teacher observations, and others

Teachers: Degrees(s), experience, in-service training, degree-granting institution(s), individual demographics (e.g., age, gender, race/ethnicity, etc.), attendance, observation ratings, evaluation ratings, student performance (e.g., grades, promotion/retention, attendance, discipline, assessment scores, etc.), student factors (see above), school factors (see below), and others

Schools (Includes staff other than teachers, districts, programs, and other groupings of students): Expenditures, resources, building factors (e.g., age, size, condition, features, etc.), staffing, staff characteristics (same as teachers above), student factors (see above), teacher factors (see above)

ESP Solutions Group

Group Measures				Individual Measures			
Pre-Measure [Acceptable Metrics]	**Post-Measure [Acceptable Metrics]**	**Calculation [Statistical Model(s) for V-A]**	**Statistical Significance Test**	**Pre-Measure [Acceptable Metrics]**	**Post-Measure [Acceptable Metrics]**	**Calculation [Statistical Model(s) for V-A]**	**Statistical Significance Test**
$Score_1$, $Year_1$, $Grade_1$, Cohort 1	$Score_1$, $Year_2$, $Grade_1$, Cohort 2	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure for Comparison Group)	**Growth:** (parametric, means; non-parametric, SEM) **Value Add:** (parametric, means; non-parametric, SEM)	Not Applicable	Not Applicable	Not Applicable	Not Applicable
[Cohort: Raw score, percent correct, scale score, percentile, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]				Not Applicable			
$Score_1$, $Year_1$, Grade X_1, $Cohort_1$	$Score_2$, $Year_2$, Grade X_2, $Cohort_1$	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure for Comparison Group) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, means; non-parametric, SEM) **Value Add:** (parametric, means; non-parametric, SEM)	$Score_1$, $Year_1$, $Grade_1$, $Student_1$	$Score_2$, $Year_2$, $Grade_2$, $Student_1$	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure for Comparison Students) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, variance; non-parametric, SEM) **Value Add:** (parametric, variance; non-parametric, SEM)
[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]				[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]			
($Score_2$, $Year_2$, $Grade_2$, $Cohort_1$) minus ($Score_1$, $Year_1$, $Grade_1$, $Cohort_1$)	($Score_3$, $Year_3$, $Grade_3$, $Cohort_1$) minus ($Score_2$, $Year_2$, $Grade_2$, $Cohort_1$)	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure Minus Pre-Measure for Comparison Group) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, means; non-parametric, SEM) **Value Add:** (parametric, means; non-parametric, SEM)	($Score_2$, $Year_2$, $Grade_2$, $Student_1$) minus ($Score_1$, $Year_1$, $Grade_1$, $Student_1$)	($Score_3$, $Year_3$, $Grade_3$, $Student_1$) minus ($Score_2$, $Year_2$, $Grade_2$, $Student_1$)	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure Minus Pre-Measure for Comparison Students) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, variance; non-parametric, SEM) **Value Add:** (parametric, variance; non-parametric, SEM)
[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]				[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]			
$Scores_{1:n}$, $Years_{1:n}$, $Grade_{1:n}$, $Cohort_1$	Target Score minus Predicted Score	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure Minus Pre-Measure for Comparison Group) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, means; non-parametric, SEM) **Value Add:** (parametric, means; non-parametric, SEM)	$Scores_{1:n}$, $Years_{1:n}$, $Grade_{1:n}$, $Student_1$	Target Score minus Predicted Score	**Growth:** Post-Measure Minus Pre-Measure **Value Add:** (Post-Measure Minus Pre-Measure) Minus (Post-Measure Minus Pre-Measure for Comparison Students) [Regression, Hierarchical Linear Model]	**Growth:** (parametric, variance; non-parametric, SEM) **Value Add:** (parametric, variance; non-parametric, SEM)
[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]				[Cohort: Raw score, percent correct, scale score, percentile, percentile growth, grade equivalent, lexile, quantile, standard score, NCE, performance level, met standard] [Value-Add: Scale score, standard score, NCE]			

References

Betebenner, Damian W. (2008). *Norm- and Criterion-Referenced Student Growth.* Dover, NH: National Center for the Improvement of Educational Assessment.

Betebenner, D. & Doran, H. (2004). A Proposal for Modeling Student Growth as Outlined by HB 04-1433. School of Education, University of Colorado, Boulder, CO. Council of Chief State School Officers, Washington, DC.

Braun, Henri I. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Princeton, NJ: Educational Testing Service.

DePascale, Charles A. (2006). *Measuring Growth with the MCAS Tests: A consideration of vertical scales and standards.* National Center for the Improvement of Educational Assessment for the Massachusetts Department of Education.

Hull, James (2008). *Growth Models: A Guide for Informed Decision Making.* Center for Public Education, National School Boards Association.

Ligon, Glynn D. (2003). *Learning Growth Index for SARs.* Austin, TX: ESP Solutions Group.

Ligon, Glynn D. (2006). *Creating a Balanced Perspective on Growth.* Austin, TX: ESP Solutions Group.

Ligon, Glynn D. (2006). *Peer Review Guidance for the NCLB Growth Model Pilot Applications.* U.S. Department of Education.

Lissitz, Robert (2005). *Value Added Models in Education: Theory and Practice.* Maple Grove, MN: JAM Press.

Rand Corporation (2004). *The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness.* Santa Monica, CA.

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

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?

ESP Insight
Growth is more for the policy makers than for the teachers.

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?

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

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.

ESP Insight
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).

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

By the way, this is all good. 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).

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.

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

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 50th percentile student; however, a lower percentile may be used to include all students who might have scored at the 50th 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

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 4th 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 7th grade. (Go ahead folks if you want your star student to suddenly be average.) Imagine what the parent of a 10th 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 50th percentile must be the grade and month representing the critical norming date (the middle date during the national norming period).
- A 50th percentile student must gain 1.0 GE a year to maintain the 50th percentile ranking.
- Students above the 50th percentile must gain more than 1.0 GE to maintain the same percentile.
- Students below the 50th 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 65th percentile is a failing grade.
- A student at the 50th percentile two years in a row may be thought not to have grown any.
- A student at the 25th percentile two years in a row may be thought to have kept pace with the average students.
- A student moving from the 12th to the 14th percentile may be thought to be catching up with the average student.
- A student scoring at the 65th percentile in both reading and mathematics may be thought to be equally above grade level in both areas.

ESP Insight

Scale scores work—if they are equal interval. If scale scores are also a vertical scale, even better.

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 99th 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).

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

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.

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 50th 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.

ESP Insight
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.

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.

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

> **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 50th 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.

ESP Insight
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.

ESP Insight
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.

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 50th 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 50th percentile)
- Making the same growth as other students with the same prior score (except for the 50th 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.

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

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 50th 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.)

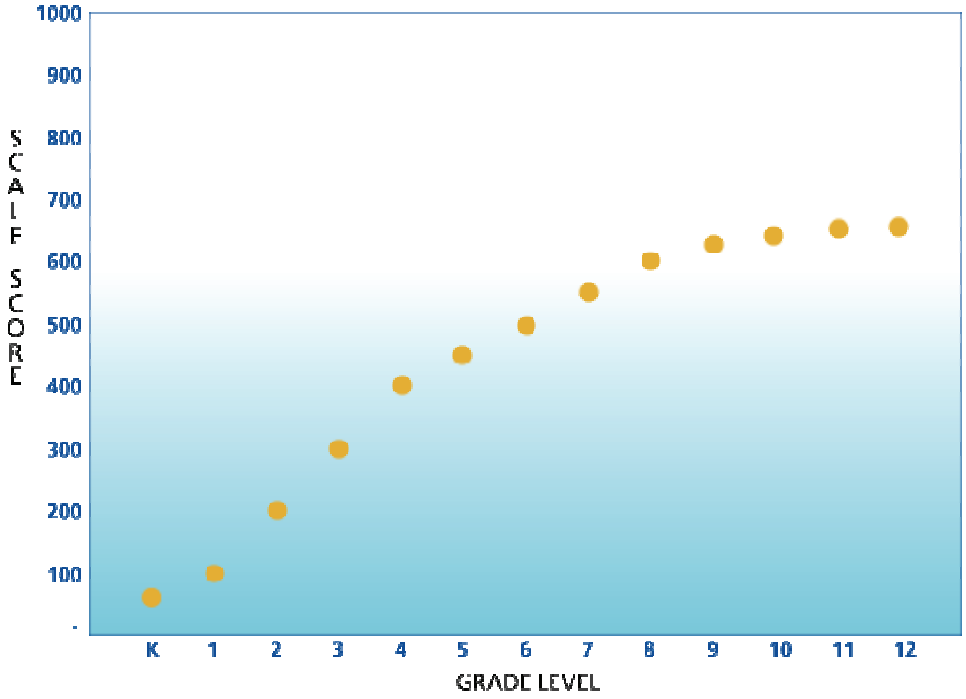
Artifactual growth can be any one of these. (Artifactual growth is further illustrated by Colorado's student growth percentile model, which is discussed later.)

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.

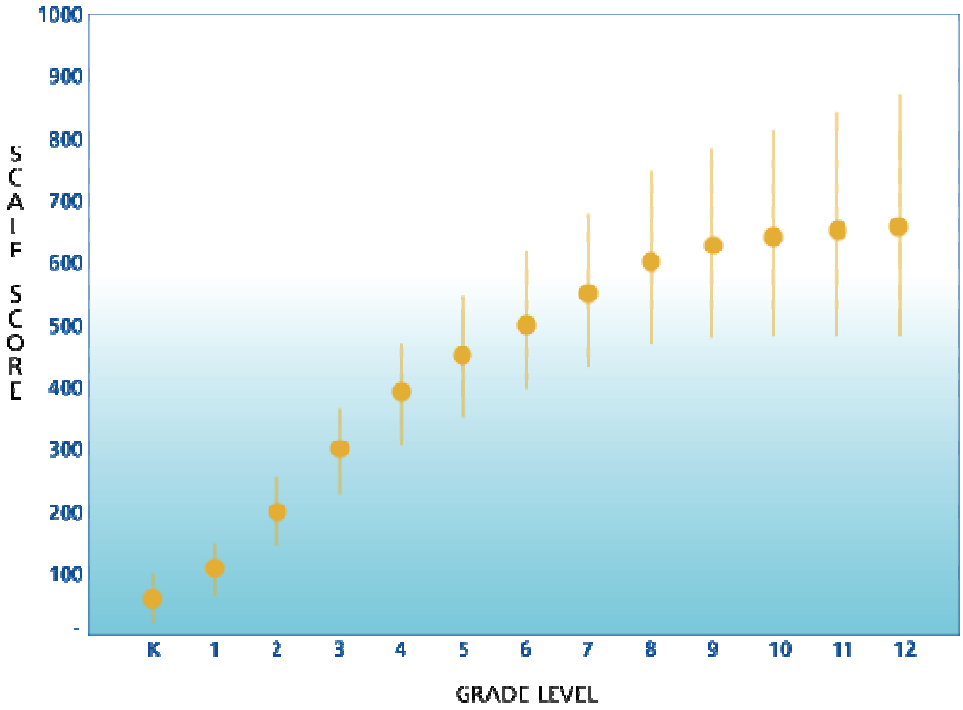
Why is this important? Growth models that use only past scores for students typically predict higher future performance than models that use real scores for higher grade levels.



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.

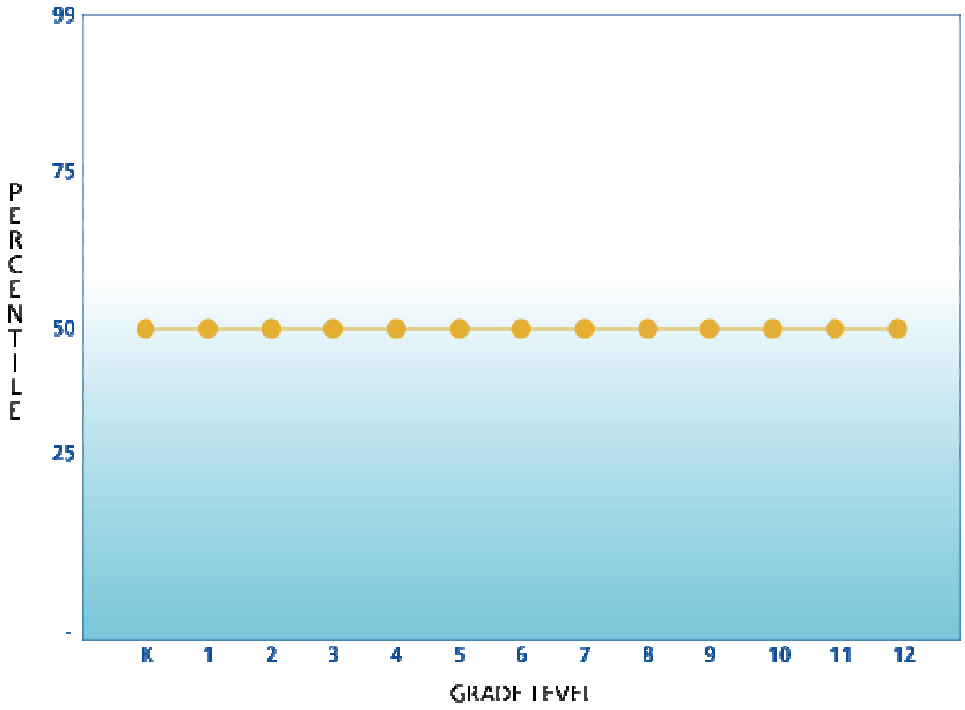
Why is this important? Growth models that predict performance are looking into a future where the variance, in this case the error, of those predictions gets higher with each grade level.



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 50th 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.

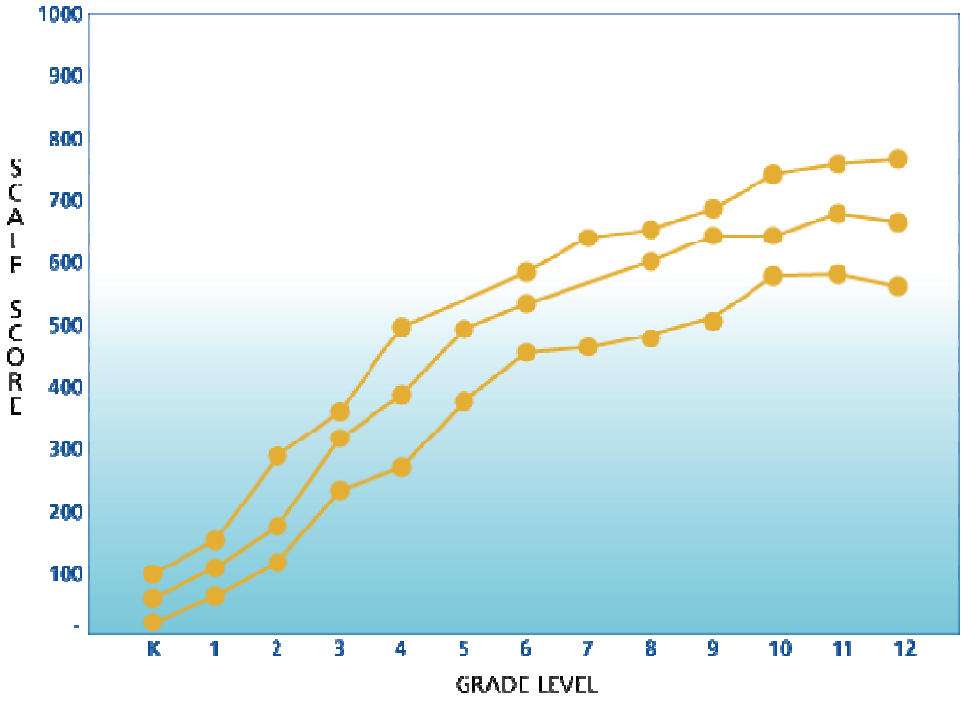
Why is this important? The average student at each grade level provides a context, a reference, a benchmark that is useful for interpreting academic performance.



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 sores.
- Variance (standard deviation) reflects this wobble and error, and results in our becoming less precise in our measurements.

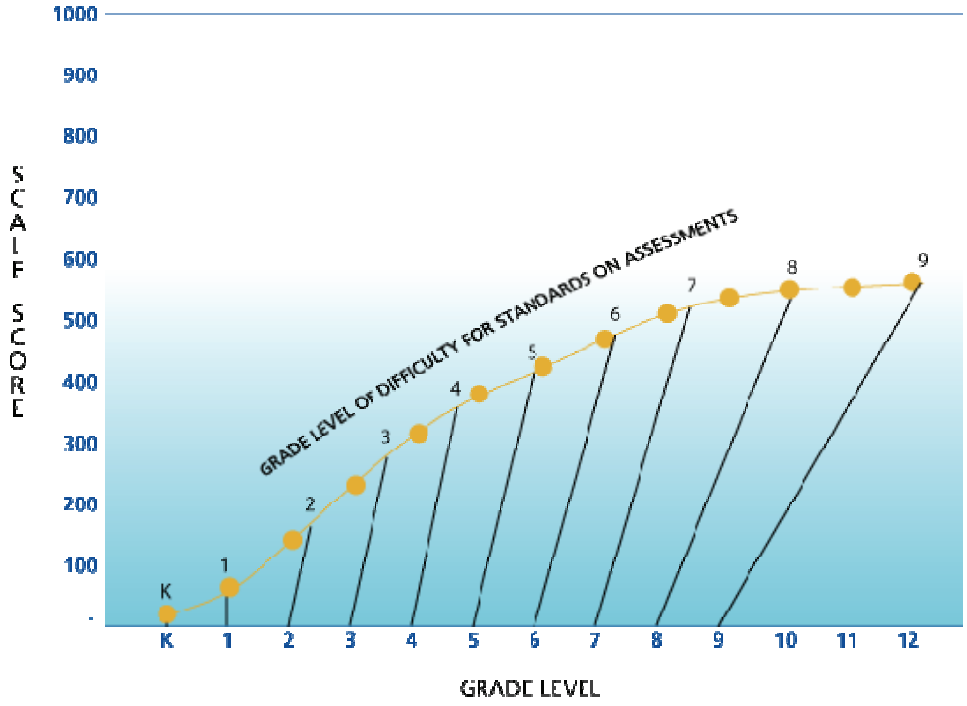
Why is this important? We probably place too much confidence in one year's test score—definitely too much confidence in a growth measure based upon more than one score.



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 11th and 12th grade levels on state assessments may be equivalent only to the average performance measured at the 9th grade level.

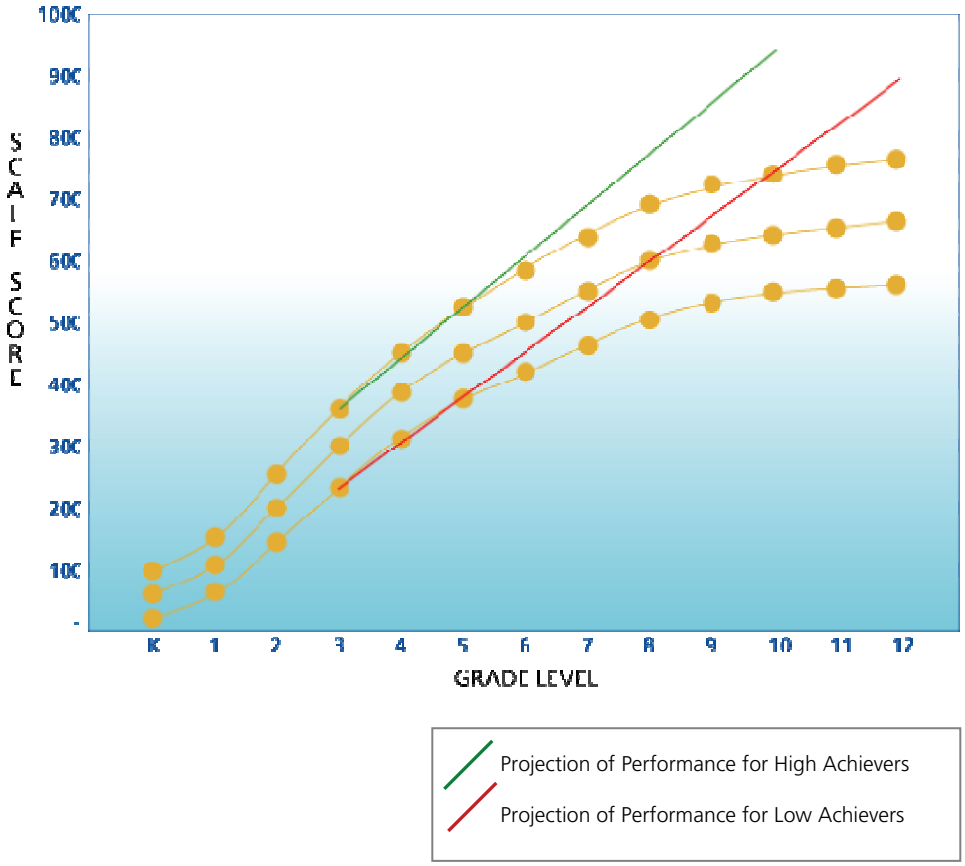
Why is this important? When we interpret performance on tests, we should understand that being on grade level or being proficient is less of a rigorous standard in secondary schools.



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.

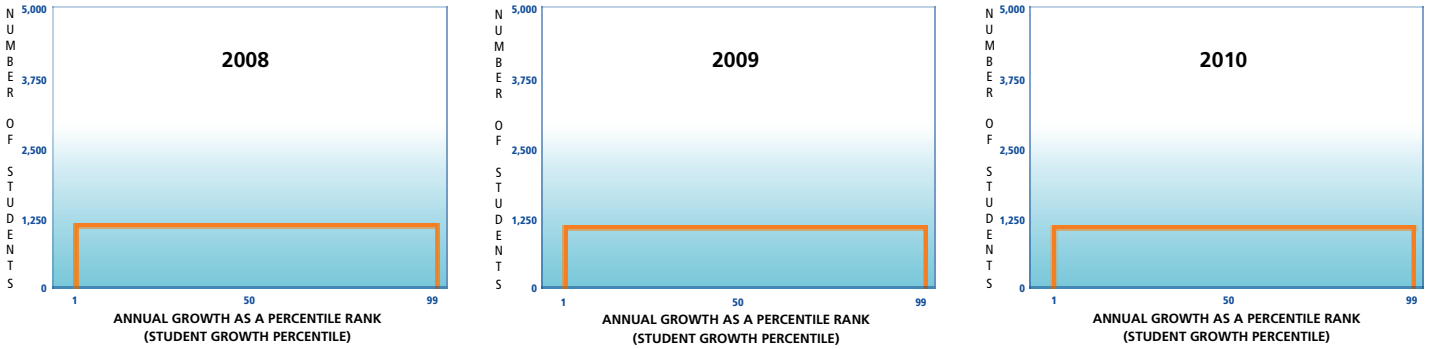
Why is this important? We mislead ourselves and others when we project future performance that is too high. At-risk students who need extra support and intervention are falsely classified as safe.



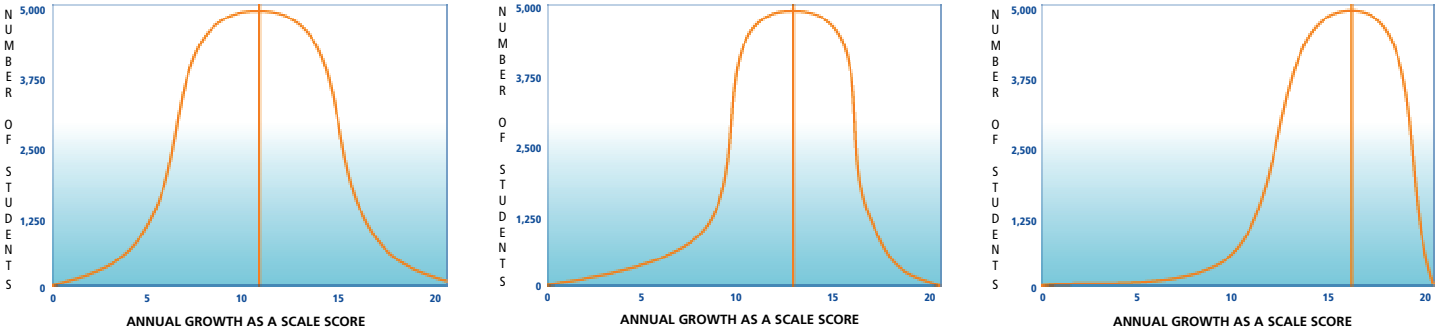
Observation 7. Baselines rule.

- 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.

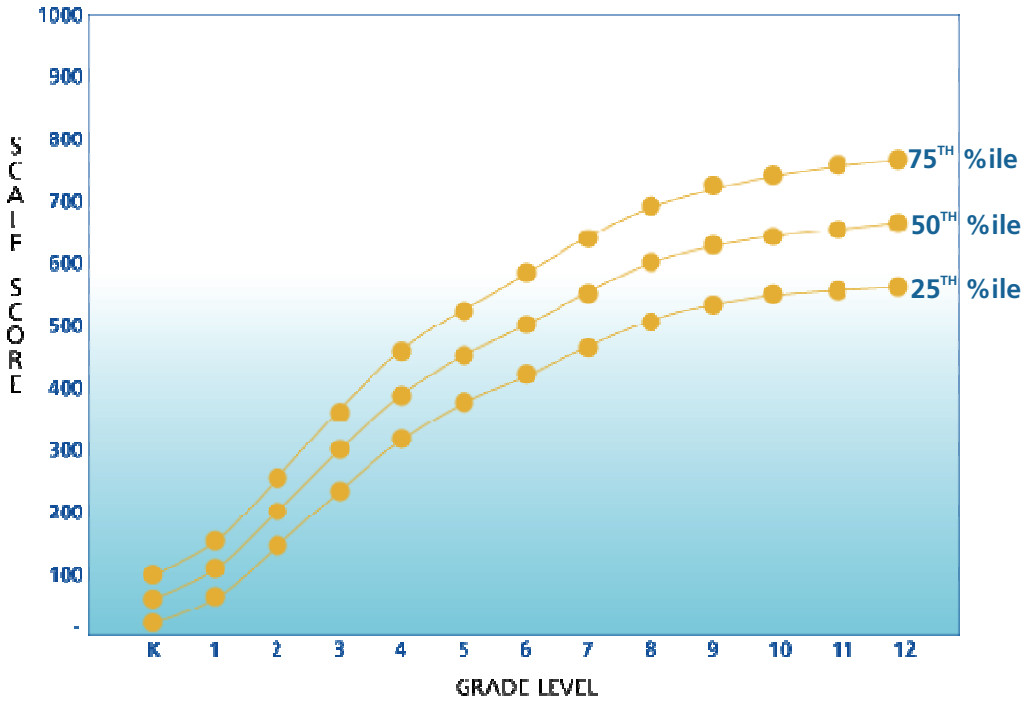


Any or all future students can outperform student in a baseline year.

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), 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 50th percentile student; however, a lower percentile may be used to include all students who might have scored at the 50th 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 50th percentile student.

About the Author

Glynn D. Ligon, Ph.D., President and CEO

Dr. Ligon, the president and chief executive officer of ESP Solutions Group, is a nationally recognized expert on issues relating to student record collection and exchange, data quality, data reporting, and large-scale system design.

The National Center for Education Statistics, the U. S. Department of Education and over 25 state education agencies have consulted with Dr. Ligon on various areas of his expertise. He has a Ph.D. in Educational Psychology, Quantitative Methods from The University of Texas at Austin and is licensed to teach in the State of Texas.

Prior to starting ESP in 1993, Dr. Ligon directed the Austin (TX) Independent School District's information and technology organization. As the executive director of management information, he led the district's efforts in developing and reporting on district-wide program evaluations, many of which won national awards from the American Educational Research Association. Dr. Ligon was also a leader in the advent of SPEEDE/ExPRESS, the EDI standard for the exchange of electronic student transcripts.

From 1992 to 2000, he served as a member of the U.S. Department of Education's Planning and Evaluation Services Review Panel. Dr. Ligon's whitepapers; *A Technology Framework for NCLB Success* and *Steps for Ensuring Data Quality* are prominently featured within the U.S. Department of Education's 2005 National Education Technology Plan, meant to help motivate and incite technology-driven transformation in education.

At the beginning of his career, Dr. Ligon taught in predominantly Spanish-speaking schools near the Texas-Mexico border. He is an experienced evaluator of Title I, Migrant, compensatory education, and bilingual education programs.

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)**, **ED*Facts***, 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.

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

ESP Optimal Reference Guides and Optimal Reference Books

ESP covers a wide variety of education topics with our series of informational whitepapers called Optimal Reference Guides (ORGs) and Optimal Reference Books (ORBs). All are available for free download at *www.espsolutionsgroup.com/resources.php*. You can also subscribe to our monthly newsletter to have ORGs and ORBs emailed to you as soon as they are published. Just visit the link above for more information.

Data Quality
- The Data Quality Imperative, Data Quality Series—Part I
- The Data Quality Manual, Data Quality Series—Part II

Data Management
- Actions Speak Louder than Data
- From Information to Insight—The Point of Indicators
- Aligning Indicators and Actions
- Data Management Strategy for States and Districts
- Defining Data
- Management of a Education Information System
- Our Vision for D3M
- Using Assessment Results to Get Performance Results
- Why Eva Baker Doesn't Seem to Understand Accountability—The Politimetrics of Accountability

Longitudinal Data Systems
- D3M Framework for Building a Longitudinal Data System
- The Dash between PK and 20: A Roadmap for PK-20 Longitudinal Data Systems
- What's Really "In Store" for Your Data Warehouse? Data Warehouse Series—Part I
- What's Behind Your Data Warehouse, Data Warehouse Series—Part II
- Accessing Student Records in a State Longitudinal Database, Data Warehouse Series—Part III

Project Management
- Why 70% of Government IT Projects Fail, Project Management Series—Part I
- From Risk to Reward: A Guide to Risk Management, Project Management Series—Part II
- Marketing Your Field of Dreams, Project Management Series—Part III

Electronic Transcripts
- Electronic Student Records and Transcripts: The SEA Imperative
- Why Your State Needs a PK-20 Electronic Record/Transcript System

Standards
- Articulating the Case for Course Numbers
- Confidentiality and Reliability Rules for Reporting Education Data
- FERPA: Catch 1 through 22
- Graduation Rates: Failing Schools or Failing Formulas?
- National Education Data Standardization Efforts
- Racial/Ethnic Data Reporting in Education
- Recommended Data Elements for EDEN Reporting
- Revisions to FERPA Guidance

Trends in Education
- Data-Driven Decision Making 2016
- How Education Information Fared in the Last Decade
- IT Defined…for the Educator
- Why My Space Matters to the K-12 Space

Student/Staff Identifiers
- Requirements for an RFP for Student Identifiers
- Statewide Student Identifier Systems

Disaster Prevention & Recovery
- Disaster Prevention and Recovery for School System Technology

Growth Models
- Growth Model Growing Pains, Growth Model Series—Part I
- Comparison of Growth and Value-Add Models, Growth Model Series—Part II
- Making a Year's Growth and Performing on Grade Level: Muddled Definitions and Expectations, Growth Model Series—Part III
- Growth Models—Finding Real Gains

ESP Solutions Group

(512) 879-5300

www.espsolutionsgroup.com