

*The Optimal Reference Guide:*

# **Growth Model Growing Pains**

## **Growth Model Series – Part I**

---

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

By Glynn D. Ligon



# **ESP Solutions Group**



## Table of Contents

Foreword.....	3
Introduction .....	4
Part 1: Do Growth and Value-Add Models Work? .....	7
The National Pastime Model .....	7
Growth Models are Not Quantum Mechanics .....	8
Why Projections are Tough .....	9
The Experts Themselves Doubt Value-Add Models.....	12
Why Pay-for-Performance Doesn't Work in Public Education.....	12
The Best Way to Determine Dollars for Performance .....	14
Other Things That Need to Change.....	17
The Socio-Economic Mandate.....	17
Statisticians Make Growth Models Too Difficult.....	18
Is NCLB Getting Growth Right?.....	18
Growth Models Won't Save Many Low-Performing Schools.....	18
We Want Our Models to Fail—Eventually.....	19
Questions for Your Growth Model Expert .....	19
Thinking Inside the Box.....	20
Limitations on Growth and Value-Add Models.....	21
The Question Parents Should be Asking .....	23
How Many Growth and Value-Add Models are There? .....	23
Awarding the Proper Value for Value-Add .....	25
Tiered Performance Indicator.....	27
Conclusions.....	27



## Foreword

By: Glynn D. Ligon, Ph.D., President and CEO  
ESP Solutions Group, Inc.

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 two-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.



- 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

**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 two Optimal Reference Guides. The two papers discuss growth and value-add models from very different perspectives. Part 1 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 2 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.



## Part 1: 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.

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.

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.



**ESP Insight**

*Baseball owners discovered that they have been using “official statistics” to judge players, but those statistics don’t relate to winning games.*



**ESP Insight**

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

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



ESP Insight

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

### 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.

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.

#### 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.

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

**Table 1: Definition of Terms**

<b>Term</b>	<b>Definition</b>	<b>Example</b>
<b>Measure</b>	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
<b>Score</b>	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
<b>Status</b>	A score on a measure at a point in time	April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score and Proficiency Level
<b>Growth</b>	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 <b>MINUS</b> April 18, 2007, Grade 4, State-Adopted Assessment, Mathematics Scale Score
<b>Value Add</b>	<b>Unadjusted:</b> 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 <b>MINUS</b> April 18, 2007, Grade 4, State-Adopted Assessment, Mathematics Scale Score <b>MINUS</b> Average difference for similar teachers, schools, or programs
	<b>Adjusted:</b> An increase or decrease made to the unadjusted value-add measure to take out the effect of selected factors	Above difference <b>MINUS</b> Average difference for similar students, e.g., race/ethnicity, gender, family income, English proficiency, etc.

<b>Projection</b>	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 <b>AND</b> April 16, 2008, Grade 5, State-Adopted Assessment, Mathematics Scale Score <b>TO</b> End of Grade 4
<b>Factor</b>	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.
<b>Weight</b>	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
<b>Model</b>	The formula that relates factors to the outcome measure	
<b>Pay-for-Performance</b>	A bonus compensation plan that uses outcome measures to determine the dollar amount of the bonus	
<b>Regression, Hierarchical Linear Model, Mixed-Effects Model, etc.</b>	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.

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.



*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.*



*To a surprising degree, private industry managers and executives just don't understand the significance of "public" in public education.*

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.

### 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?



*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.

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.



**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.*

**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.

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.” (*“Why Eva Baker Doesn't Seem to Understand Accountability—The Politimetrics of Accountability,” ESP Optimal Reference Guide, 2008. Available for download at [www.espsg.com/resources.php](http://www.espsg.com/resources.php).)*



*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.*

## 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:

- 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.



*The fact is, everything doesn't matter.*

## 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 ("*Why Eva Baker Doesn't Seem to Understand Accountability—The Politimetrics of Accountability*," *ESP Optimal Reference Guide, 2008*. Available for download at [www.espsg.com/resources.php](http://www.espsg.com/resources.php).), 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.

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

## 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.



**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.*

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?
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.



**ESP Insight**

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

## **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 outweigh the benefits of having growth and value-add information in our D3M processes.

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.
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.

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

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 2 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 2 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.



*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.*

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 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.



**ESP Insight**

*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.*

### **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

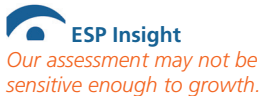
celebration and recognition. However, the school is below the goal for student performance, and the school's growth trend is too slow to reach the goal by a reasonable target date.

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.

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 meeting 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.



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

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.

This methodology will be expanded in Part 2

## 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.



### 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)**, **Education Data Exchange Network (EDEN)**, and the **Schools Interoperability Framework (SIF)**.

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

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

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

*Growth Model Growing Pains, Growth Model Series – Part I.* Copyright © 2008 by ESP Solutions Group. All rights reserved. No part of this paper shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher.



# ESP Solutions Group

(512) 879-5300

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