Growth Model Comparison Study:
A Summary of Results

A summary for practitioners based on the multi-state *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance*

conducted by
Pete Goldschmidt, Kilchan Choi, and J.P. Beaudoin

Prepared by:
Frank Brockmann and Bill Auty

for the
Technical Issues in Large Scale Assessment (TILSA) and
Accountability Systems & Reporting (ASR)
State Collaboratives on Assessment and Student Standards (SCASS)
of the Council of Chief State School Officers

Growth Model Comparison Study: A Summary of Results

A paper commissioned by the
Technical Issues in Large-Scale Assessment
and Accountability Systems & Reporting
State Collaboratives on Assessment and Student Standards
Council of Chief State School Officers

Authored By:
Bill Auty, Education Measurement Consulting
Frank Brockmann, Center Point Assessment Solutions

Supported By:
Charlene Tucker, TILSA Advisor
Duncan MacQuarrie, Associate TILSA Advisor
Doug Rindone, Associate TILSA Advisor

Based on Research and Commentary From:
Pete Goldschmidt
Kilchan Choi
J.P. Beaudoin

Special Thanks:
Arie van der Ploeg, American Institutes for Research

This report was prepared for the Technical Issues in Large Scale Assessment (TILSA) and Accountability Systems & Reporting (ASR) members of the system of State Collaboratives on Assessment and Student Standards (SCASS) supported by the Council of Chief StateSchool Officers (CCSSO).

Contents

The Purpose of this Publication

School accountability is subject to considerable scrutiny. It generates sharp political debate, policy challenges, and continuous discussion. Growth models are now a part of that discussion. To many practitioners the sheer volume of "important to know" information is daunting.

The members of the Technical Issues in Large Scale Assessment (TILSA) and Accountability Systems & Reporting (ASR) state collaboratives[1] have recognized the challenge of communicating detailed, technically-oriented measurement issues. In response, these groups have produced a series of publications aimed at providing technical and non-technical users with practical guidance about growth models.

This publication continues the series by summarizing the *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance*[2] conducted by Pete Goldschmidt, Kilchan Choi, and J.P. Beaudoin which implemented statistical growth models using real student data from four states. The study is important because it is the first designed to make it possible to see whether results are different or similar when (a) the same growth models are estimated in different states and (b) different models are estimated in the same state.

As a summary, this document is intended to provide timely and comprehensible information for practitioners and other stakeholders who may not have a technical background in using assessment data for educational accountability.

1 These two state collaboratives operate under the aegis of the Council of Chief State School Officers (CCSSO).
2 Pete Goldschmidt, et al., (Washington, DC: CCSSO, 2012).

Executive Summary

Many states have begun to use growth models for school accountability. These models frequently differ in their technical details. Goldschmidt et al.'s *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance* used real student assessment data from four states to compare various growth models and analyzed how they performed against each other and how they performed over time. The study sought to answer questions practitioners and policymakers ask:

- Overall, does the model matter?
- Do different models lead to different inferences about schools?
- How accurately do models classify schools into performance categories?
- Are models consistent in classifying schools from one year to the next?
- How are models influenced by school intake characteristics (percent ELL, FRL[3], etc.)?
- Do models perform similarly for elementary and middle schools?
- Do models behave similarly across states?

The primary conclusion of the *Growth Model Comparison Study* is that the choice of growth model matters.

Other findings include the following:

- Different models can lead to different inferences about schools. An "A" school under one growth model may be a "C" school under another.
- Models varied considerably in regard to how well they classified schools into performance categories. The variations are partly due to the construction of the models themselves, and partly due to the heterogeneity of the schools in the data set. The simpler models that only used test scores from two successive years were more influenced by these variations; more complex models that used more than two years of test scores or other background data were least influenced.
- Most of the models were fairly consistent in classifying schools from one year to the next.
- Growth models performed differently from state to state. This is a critical result. Different assessments, testing practices, and the characteristics of schools within and between states all play a role. Simply adopting a model that worked for one state for use in another state may not be a good policy option.

No model will fit all situations as each model has specific uses and implications; careful consideration will be needed to choose a growth model that fits. This document informs that consideration.

3 Refers to *English language learner* and free or *reduced lunch eligible*, respectively.

An Introduction to the *Growth Model Comparison Study*

Recent History/Background

The 2002 reauthorization of the Elementary and Secondary Education Act (ESEA) — commonly referred to as No Child Left Behind (NCLB) — changed school accountability and stiffened its requirements. Since then, more recent federal initiatives such as Race to the Top and the ESEA flexibility guidelines continue to push states to develop new ways to accurately and fairly monitor school performance.

There is now considerable agreement among educators that monitoring schools using only status indicators such as "the percentage of proficient students" is not an effective or fair way to hold schools accountable for student achievement.[4] That is, they focus more on what students bring to school than on how they change while there. This emphasis places schools with more homogeneous or fewer at-risk students at an advantage.[5] As a result, many states and districts have begun to incorporate growth measures into their examinations of school performance to track whether students' learning changes with time.

Even with a fundamental understanding of what growth models are and what their intended purpose might be, practitioners are still left with a host of pressing questions.

Even with a fundamental understanding of the variety and limits of growth models and what their intended purposes may be, practitioners are left with a host of pressing questions.

Goldschmidt et al. designed the *Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance*[6] to address these questions by using real student data from several states to compare growth models. Their study provided detailed technical analyses leading to important conclusions about how growth models perform, how they compare to each other, and the kinds of claims or inferences they might support.

The Purpose of the Study

A primary purpose of the *Growth Model Comparison Study* was to help states select growth models for school accountability; to that end, the study was designed to provide information that would give states a greater sense of "informed flexibility." The authors state their goal was "...to determine the potential latitude states might have in choosing a growth model for school accountability."

4 Goldschmidt, et al. reference Novak & Fuller, "Penalizing diverse schools? Similar test scores, but different students, bring federal sanction," *PACE Policy brief* (Berkeley, CA: Policy Analysis for California Education, 2003). See also Rothstein, Jacobsen, & Wilder, *Grading Education: Getting Accountability Right* (New York, NY: Teachers College Press, 2008).
5 Ibid.
6 Goldschmidt, et al., *Growth Model Comparison Study*, 5.

In addition to the detailed analysis of how the various growth models compare, the researchers also hoped to identify what the potential tradeoffs may be in using one model versus another. In this way, the study was intended to serve as a basis for practical recommendations regarding the use of growth models for school accountability, and as practical guidance in what to consider when evaluating how well a growth model is meeting its intended goals. By detailing the results of the comparative analysis, Goldschmidt et al. hoped to give states a clearer and more empirically-based guide to their options.

Note that the focus of the study was the use of growth models for school accountability in general. Although it provided some information about other uses, the study did not comprehensively explore the use of growth models for any single specific purpose such as teacher evaluation.

The Researchers' Approach

Most elements of the design of assessment and accountability systems depend on *purpose*. If states are unclear about what they are using growth model data for and what inferences or claims they are looking to support, growth model results, regardless of the model, will be difficult to interpret and apply.

In designing the *Growth Model Comparison Study*, the researchers began with a set of assumptions about growth models and their role with regard to accountability:

> *"We assume that the primary impetus for using growth models is to correctly identify the spectrum of school effectiveness in order to accurately monitor schools…This focus is somewhat different from a system that provides information on which schools attain the highest achievement levels without consideration of potential confounding factors*[7]*…Hence, our concern is to first provide policymakers with results that afford valid inferences related to school performance."* [8]

Goldschmidt et al. identified several questions they wanted to answer. These questions are presented below, along with brief commentary:

Overall, does the model matter?

> This is an important question because previous studies that compared growth models did not typically have the scope or breadth of the *Growth Model Comparison Study*. Generally, the prior research was more limited; it used simulated data and focused on very specific aspects of the models, it included data from different data sets, but did not make comparisons across states, and/or it did not compare growth across cohorts. Therefore, it is sensible to ask whether different models will produce different results when applied to the same data set—one which contains real student results from multiple states and cohorts. We can then answer the question by stepping systematically through a series of carefully documented models.

7 One criticism of achievement-level-oriented accountability systems is that there is not enough attention paid to other factors, such as the prior achievement history of the students in the school, proportion of students coming from families below the poverty level, and the percentage of special needs students enrolled.
8 Goldschmidt, et al., *Growth Model Comparison Study*, 4.

Do different models lead to different inferences about schools?

If so, the inferences may be driven by the structure of the models or by data from the schools or systems under evaluation. It is critical to be certain that the mechanism for interpretation (the model applied by the analysis) is responding to the questions to be answered. To infer that school A is doing better than school B, the model must support such an inference and be explicitly clear about the standard of comparison.

How accurately do models classify schools into performance categories?

Federal and state accountability systems are typically designed to place schools into performance categories, that is, to "grade" schools (A, B, C, D, and F). In a sense, this is just a technical exercise. But it is critical to understand how differences in growth models affect the grouping of schools into these categories.

Are models consistent in classifying schools from one year to the next?

Growth models assess the growth of a group of students whose performance is tracked over time. Models need to be sensitive to changes in student performance (and related characteristics), but they also need to demonstrate some *stability*. For example, if many schools are given a "D" rating one year and an "A" rating the following year, these results erode confidence in the rating system. We must be confident that we are measuring growth consistently from one year to the next, and that the growth we are measuring reflects school processes.

How are models influenced by school intake characteristics[9] (percent ELL, FRL, etc.)?

The issue of including or excluding student background characteristics in accountability models has both statistical and philosophical ramifications. However, the important question for the *Growth Model Comparison Study* was whether they matter *statistically*.

Do models perform similarly for elementary and middle schools?

If models do not operate consistently across levels of schooling, this suggests either a weakness in the model or a distinction in the educational processes enacted. It will be important to know which is the case if differences are found.

Do models behave similarly across states?

States have developed assessment systems that are independent of each other, and the characteristics of those assessments may cause growth models to work differently. Similarly, it is possible that the nature and distribution of schools and students between states is meaningfully different. In such cases we would not necessarily expect growth models to behave similarly.

9 The term "school intake characteristics" generally refers to demographic factors known to be associated with school outcomes, such as school composition, urban/rural status, socioeconomic status, participation in free and reduced lunch programs, percentage of English language learners, and the like. ELL and FRL refer to *English language learner* and *free or reduced lunch eligible*, respectively

Challenges in Defining and Classifying Growth Models

The issue of how best to define and classify growth models is always open to debate and discussion, and there is no current encyclopedic reference that perfectly defines and categorizes every growth model researchers might want to consider for every situation.

In A Practitioner's Guide to Growth Models,[10] Castellano and Ho offer guidance about growth model use, classification, and interpretation and provide the following definition:

> **A *growth model* is a collection of definitions, calculations, or rules that summarizes student performance over two or more time points and supports decisions about students, their classrooms, and their educators, or their schools.**

Castellano and Ho present critical questions for growth models and answer these questions for a range of models used in practice. One critical question concerns the primary interpretation that each growth model supports:

- Models that focus on describing growth
- Models that attempt to predict growth in the future
- Models that attempt to identify the causes of growth

In the *Growth Model Comparison Study*, Goldschmidt et al. classify growth models by focusing on more specific interpretations that their calculations support[11]:

- ***Categorical Models*** use the change in student performance category placement from year to year as the growth indicator.
- ***Gain Score Models*** are based on the difference between a student's earlier score and a later score. Gains can provide a simple estimate of change, but may have low reliability.
- ***Regression Models*** can provide the most precise measure of growth. Calculations are complex and a vertical scale (see sidebar) is generally required. These models can estimate gain between two scores or a slope over more than two time points.
- ***Value-Added Models*** are a complex type of regression model that takes into account student or school characteristics. Value Added refers to producing more growth than expected, given specific characteristics of the student or school.
- ***Normative Models*** compare changes in student performance to a norm group to determine whether change is typical, high, or low. A vertical scale is not required. These models do not directly address whether the observed growth is adequate to reach a defined standard.

Growth and Vertical Scales

Ideally, growth should be calculated from test scores based on a scale that is consistent for the period of time involved. This is called a *vertical scale*, and differences in scores have a consistent meaning both within and between years.

The *Growth Model Comparison Study* cites research indicating that such a continuous scale is often not available for statewide assessments. However, other research into the estimation of growth indicates that the statistics may still be robust enough to draw important conclusions about growth. Users of growth models are cautioned to ensure that the assessment system supports developing a vertical scale if that is what the growth model requires.

10 Katherine Castellano and Andrew Ho, *A Practitioner's Guide to Growth Models* (Washington , DC: CCSSO, 2012).
11 The Goldschmidt et al. classifications are based on those found in the brochure *Achievement Growth and Accountability: What to Look For—and What to Look Out For* (Washington, DC: CCSSO, 2011).

Growth Models Used in the Study

Although many different variations of growth models exist, Goldschmidt et al. compared and contrasted nine distinct models which represent the more common varieties found in current practice:

Table 1: Summarized Descriptions of Growth Models Used in the Study

Growth Models Used in the Study		
Model / Acronym	***Type***[12]	***Summary Description***
Simple Gain (GAIN)	Gain	Uses the difference between a student's test scores from two points in time—the "simple gain." Simple gain is calculated for all students, and averages are calculated for schools.
Fixed Effects Gain (FEG)	Gain	Closely related to a simple gain model because gains are the primary outcome of interest. The difference is that a fixed effects model directly provides estimates of precision and introduces an error component into the model for additional ways to check the model's performance.
True Score Gain (TSG)	Gain	Does not use observed gains as an outcome, rather it estimates true gains as part of the model, and hence avoids the spurious negative correlation between initial status (pre-test) and gains.
Growth To Standards (GTS)	Gain	Based on gains; however, growth is defined in relation to an external standard such as the state proficiency score. Therefore it directly addresses the question, "Has the student gained enough?"
Covariate Adjusted (CAFE) with School Fixed Effects	Regression	Compares schools based on their students' current scores while considering how those same students performed in the prior year(s). If schools are treated as *the population of schools,* then they are modeled as "fixed."
Covariate Adjusted (CARE) with School Random Effects	Regression	Compares schools based on their students' current scores while considering how those same students performed in the prior year(s). If schools are treated as *a sample of schools,* they are modeled as "random."
Simple Panel Growth (PANEL)	Regression	Allows for incomplete data due to mobility and other causes. Takes the nature of data into consideration and attempts to mitigate potential confounding factors.
Layered Model (LM)	Value Added	Simultaneously models scores for multiple years in multiple subjects. Later years of teacher or school effects build on earlier years. The model handles missing data well.
Student Growth Percentile (SGP) (Quantile Regression)	Normative	Compares schools based on their students' current performance while considering how those students performed in the prior year(s) in terms of student percentile ranks. The SGP explicitly compares students with the same prior ranking(s). Median percentile is calculated for each school.

12 Goldschmidt, et al., *Growth Model Comparison Study,* 15. The researchers used the classifications described in the "Challenges in Defining and Classifying Growth Models" sidebar on page 6, but also expanded upon it by presenting two ways (Intent and Estimation) to consider each of the five models. For summary purposes, Table 1 uses only the Goldschmidt et al. classifications for *intent*. In *A Practitioner's Guide to Growth Models*, Castellano and Ho identify "statistical foundations" for each growth model. These align closely with the growth model "Type" in Table 1.

About the Data

Goldschmidt et al. compiled four years' worth of large scale assessment data for approximately 675,000 students enrolled in 2,645 public schools from four participating states: Delaware, Hawaii, North Carolina, and Wisconsin. The study included individual student test results from two cohorts of students and utilized three years' worth of test results for each cohort group, as illustrated below.[13]

Year	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
2007	Cohort 1			Cohort 3		
2008	Cohort 2	Cohort 1		Cohort 4	Cohort 3	
2009		Cohort 2	Cohort 1		Cohort 4	Cohort 3
2010			Cohort 2			Cohort 4

Note: the Growth Model Comparison Study utilized cohort groups from both elementary school and middle school. Cohorts 1 and 3 represent data sets from 2007 to 2009, and Cohorts 2 and 4 represent data sets from 2008 to 2010.

The following student characteristics were collected for each student each year of the study: minority status, economically disadvantaged (ED or FRL – eligible for Free or Reduced Lunch), ELL (English language learner), SWD (students with disabilities), and mobility. Also collected were school type (elementary or middle) and school enrollment.

The Results

Primary Conclusion: Yes, the Model Does Matter

Question: Overall, does the model matter?

Yes. The study confirms that there is no single model that is better than the rest in all situations and circumstances. Goldschmidt et al. note there is no "best" model for two reasons: first, different models address different questions about schools; second, the results show the state in which the model will be run affects how the model may work. [14]

In other words, *context matters*. Test scales, testing procedures, student characteristics, and school characteristics can all play a part in producing different results from state to state, even when the same model is used.[15] Thus, there can be no single "best" growth model when the context for using *any* model can vary so significantly and when there are different options for how to address that context.

13 Goldschmidt et al. credit previous studies for providing significant guidance for the models and specifications used in their Growth Model Comparison Study.
14 Goldschmidt, et al., 54.
15 Ibid.

Goldschmidt et al. also assert that an accountability model should not be unduly influenced by factors outside of schools' control, and the growth models in the study clearly differ in this respect:

"Distinguishing between a school's ability to facilitate learning and a school's performance as a function of advantageous (or challenging) student enrollment characteristics is where statistical machinery provides its biggest benefit."[16]

No single [growth] model can unequivocally be assumed to provide the best results.

However, some are clearly better than others in terms of being able to attribute student learning to schools.

To that end, the **Growth Model Comparison Study** supports this claim by showing that some models—especially those that make use of additional information such as student background, prior performance, or multiple assessments—are less susceptible to influence beyond school control (e.g., a school's student intake characteristics). [17]

While accountability models based on growth provide a significant step toward attributing learning to schools, causal claims may still not be warranted. And, the study's findings do not mean the more complex models are always the most appropriate choice. According to Goldschmidt et al., the tradeoff is technical complexity and true understanding of the inferences afforded by the model's results.[18] And although the primary conclusion of the study is *the model does matter*, this does not mean that different models will produce entirely different results. That is, the most accurate real-world answer might be "Well, it depends…." Whether or not the model matters for any given situation depends on a variety of factors — factors which the authors address by responding to the other questions the study sought to answer.

Classifying Schools: Similar and Dissimilar Inferences

Question: Do different models lead to different inferences about schools?

It is unlikely that one model would rate a school as a top performer while another would rate the same school as a poor or very poor performer.

If the first conclusion of the study is, "Yes, the model does matter," the next conclusion might be illustrated by adding the phrase "and *one* way it matters is that different models can lead to different inferences about schools."

Goldschmidt et al. demonstrate this concept by stating that we might end up classifying a school as an "A" under one model and a "C" under

16 Goldschmidt, et al., *Growth Model Comparison Study*, 54.
17 Ibid.
18 Ibid.

another model if we use two models that are fundamentally different.[19] In practical terms, however, the models in the study showed a reasonable amount of consistency in this regard, and none of them showed extreme variations in school ratings.[20]

The Accuracy of School Classifications

Question: How accurately do models classify schools into performance categories?

The study also addressed how *accurately* models classify schools into performance categories. What the researchers found was that models vary considerably in this regard.[21] They attribute part of the variation to the models themselves, as well as the school size and something researchers call the ICC or "intraclass correlation" (a measure of how homogeneous students are within a school).[22]

The Year-to-Year Consistency of School Classifications

Question: Are models consistent in classifying schools from one year to the next?

This is an important consideration for policymakers and practitioners because unstable results would lead to a loss of credibility [of the accountability system].[23] A school's performance in one year, all else equal, ought to be a good indicator of its performance for the following year.[24]

Overall, the models show moderate consistency in classifying schools from one year to the next.[25] The least stable model in the study was the Simple Gain (GAIN) model, with other models showing varying levels of relative stability.[26] (Generally, the SGP and covariate adjustment models—CARE and CAFE—tend to be more stable than gain-based models such as GAIN, TSG, FEG, and GTS.) In general, elementary school results tend to be more stable than middle school results.[27] Put another way, the more data in the model, the more robust and therefore consistent it is likely to be.

Another concern practitioners may face is the impact of school size; the study found that some growth models are more sensitive to the amount of data available. For example, GAIN and SGP models are substantially less stable when schools are small, whereas results for the PANEL model are largely unaffected down to school sizes of 30.[28]

Some models are substantially less stable for small school sizes.

In general, elementary school results tend to be more stable than middle school results.

19 Goldschmidt, et al., *Growth Model Comparison Study*, 40.
20 Ibid., 41.
21 Ibid.
22 Ibid., 33.
23 Ibid., 56.
24 Ibid., 46.
25 Ibid., 48.
26 Ibid., 47.
27 Ibid.
28 Ibid., 48.

The Effects of School Intake Characteristics

Question: How are models influenced by school intake characteristics[29] *(percent ELL, FRL, etc.)?*

In the *Growth Model Comparison Study,* Goldschmidt et al. considered the influence of school intake (demographic) characteristics on growth model results to be a critical part of the analysis—especially because the findings provide some evidence related to bias and fairness.[30]

The results of the study indicate there is variability between models with regard to the influence of intake characteristics on the models' results.[31] For example, the researchers conclude that models most closely related to status[32] (such as GTS) tend to be *most* influenced by intake characteristics, and much of the stability observed with these models is based on the stability of school enrollments with respect to these characteristics. Models that incorporated multiple assessment results (SGP, PANEL, and LM) were the *least* influenced by student background.

Models that incorporated multiple assessment results were the least influenced by student background.

Goldschmidt et al. take a cautionary tone in describing what these findings mean for practitioners:

> *"To the extent that uncontrollable school input characteristics influence results, the model results need to be carefully examined."*[33]

The issues around the inclusion of factors related to growth, but outside the direct control of schools, are both technically and politically complex.

Elementary vs. Middle School Results

Question: Do models perform similarly for elementary and middle schools?

To address this question, the researchers compared results across models and school level to see whether they could identify specific areas where policymakers might need to use caution when applying a model.[34] They found that overall, the models in the study tend to work similarly in elementary school and middle school, although at middle school the models are sometimes less

29 As noted previously (p. 5), the term "school intake characteristics" generally refers to demographic factors known to be associated with school outcomes. ELL and FRL refer to English language learner and free or reduced lunch eligible, respectively.
30 Goldschmidt, et al., *Growth Model Comparison Study*, 55-56.
31 Ibid., 56.
32 As described by Castellano and Ho in *A Practitioner's Guide to Growth Models*, status refers to the academic performance of a student or group at *a single point in time*. In contrast, *growth* refers to as the academic performance of a student or group *over two or more time points*.
33 Goldschmidt, et al., 56.
34 Ibid., 32.

accurate and less stable.[35] The general conclusion, then, is that growth models appear to be fairly robust to school organization[36]; that is, they are likely to produce useful results for schools which test students at multiple grade levels. Conversely, Goldschmidt et al. note that K-2 schools, schools that test only one grade, or schools with untested grades (i.e., grades 9 and 10) will probably need to make different decisions about how to attribute growth.

The Behavior of Models Across States

Question: Do models behave similarly across states?

For the study, Goldschmidt et al. found that models work differently for different types of schools, and the effect varies by state. The *Growth Model Comparison Study* then discusses many factors which may be associated with these state-to-state differences.

First, because different states have different assessments, models can function differently given the characteristics of those assessments—and the results cannot be predicted completely *a priori*.[37]

Second, an obvious difference among states is the nature of the scale used for the assessments. More specifically, it depends whether or not a state has a vertical scale. The lack of a sound vertical scale is particularly problematic for the gain-based models (GAIN, FEG, TSG, and GTS).[38] States also differ with regard to testing and accountability policies. For example, one state allows students to retest for adequate yearly progress (AYP) purposes, and this practice seems to have the effect of stabilizing results.[39]

Much of the variability in how the models perform from state to state is due to differences in schools.

Most important, perhaps, is the observation that much of the variability in how the models perform from state to state may be due to differences in school populations and distributions. In the *Growth Model Comparison Study*, researchers address this variability by disaggregating data; then, they conclude that the ability of a model to distinguish school performance is likely influenced by the context.[40]

Ultimately, the study demonstrates that model results vary in substantively important ways and that no single growth model can meet the needs of every system. For example, researchers found the GTS model does quite well with advantaged schools, but substantively less well with

35 Goldschmidt, et al., *Growth Model Comparison Study*, 52.
36 Ibid.
37 Ibid., 57.
38 Ibid. For readers with technical training, Goldschmidt et al. qualify the difference by stating, "There can be 40 point swings in reliability and large swings in precision as well for a model across states." (52)
39 Ibid., 57.
40 Goldschmidt, in personal communication (1/30/2012): "These models may all work well for large, homogenous schools," he explains, "but they start to differ when schools face more challenging circumstances such as a high proportion of ELL or mobile students."

disadvantaged schools.[41] Thus, just as no single test can meet all possible assessment purposes, no single growth model is appropriate for all school accountability systems.

Concluding Commentary

Goldschmidt et al. recommend that policymakers and practitioners realize from the outset that growth models may identify some traditionally "good" schools as not performing well in terms of growth; all stakeholders *want* growth, not all schools actually *exhibit* growth—and this often comes as a shock. Goldschmidt et al. also note that including measures of growth in some way fundamentally changes how we monitor schools, and growth models are not a way to "make everyone look good."[42]

The *Growth Model Comparison Study* clearly shows that no model will fit every situation or serve every purpose. Growth models have intended uses that should be considered *before* their application, and simply importing a model to use for school accountability is not a good policy option. In short, the model matters.

The authors also emphasize the importance of understanding the fundamental nature of the data. The study identifies issues such as the presence of a vertical scale, variations in how scales are developed, intra-class correlations, and other factors that can influence how models behave.

The *Growth Model Comparison Study* suggests that policymakers and practitioners have a general sense of how much variability in growth can be attributed to schools, but this is difficult to quantify and put into policy. For example, the policy of allowing students multiple attempts in each grade is likely to stabilize results, but this policy also makes it more difficult to attribute the variation in performance to schools. If policies create very similar growth trajectories among schools, then it is difficult to distinguish among schools based on growth because there is no variation among the schools (or because the growth in all schools is equal). The types of schools a state has, the school characteristics, and school size also impact the model's results. Decisions related to stability and how to increase it (such as averaging model results over time, or using a model that includes multiple years) need to balance the tradeoff between having stable results versus having a system that is sensitive to true changes in performance.[43]

Policymakers and practitioners need to consider the tradeoff between having stable results versus having a system that is sensitive to true changes in performance.

41 Goldschmidt, et al., *Growth Model Comparison Study*, 52.
42 Goldschmidt, personal communication, 1/30/2012.
43 Goldschmidt, et al., *Growth Model Comparison Study*, 57.

Lastly, according to Goldschmidt et al., a school accountability model that aims to use both growth and status needs to carefully check the properties of each and how they work in conjunction to identify school performance. A model that uses both growth and status may unintentionally disadvantage schools that have either very high or very low test scores; in some cases the growth indicators may actually counterbalance the status indicators – resulting in many schools being deemed average, few schools being exemplary or failing[44], and creating a system in which it is virtually impossible to move out of the average range.

Simply importing a model to use for school accountability is not a good policy option. No model will fit all situations, and models have intended uses that should be considered before their application.

Even after (or during) model selection, analyses should be considered to evaluate whether models produce results that will be amenable for inferences in high stakes environments.

44 Or, for example, when using an A-F system that is being widely adopted, most schools will be a C, with few A's or F's.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072