POLICYMAKERS' GUIDE TO GROWTH MODELS FOR SCHOOL ACCOUNTABILITY: HOW DO ACCOUNTABILITY MODELS DIFFER?

A paper commissioned by the CCSSO Accountability Systems and Reporting
State Collaborative on Assessment and Student Standards

The Council of Chief State School Officers, Washington, DC

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

DIVISION OF STATE SERVICES AND TECHNICAL ASSISTANCE

The Division of State Services and Technical Assistance supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing and using information for decision-making; development of assessment resources; creation of high-quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success.

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

The State Collaborate on Assessment and Student Standards (SCASS) Project was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation.

Council of Chief State School Officers
David P. Driscoll (Massachusetts), President
Valerie Woodruff (Delaware), President-Elect
Douglas D. Christensen (Nebraska), Vice President
G. Thomas Houlihan, Executive Director

Julia Lara, Deputy Executive Director, Division of State Services and Technical Assistance
Rolf K. Blank, Director of Education Indicators

Copies of this report may be ordered for $10.00 per copy from:

Policymakers' Guide to Growth Models for School Accountability: How do Accountability Models Differ?

Pete Goldschmidt
Pat Roschewski
Kilchan Choi
William Auty
Steve Hebbler
Rolf Blank
Andra Williams

October 2005

The Council of Chief State School Officers, Washington, DC

Table of Contents

Lists of Tables and Figures

Introduction: Policymakers' Need for Information on Growth Models

The Council of Chief State School Officers (CCSSO) is working to respond to the growing interest in the use of growth models for school accountability. While growth models have been used for decades in academic research and program evaluation, now a wide cross section of policymakers at local, state, and national levels are inquiring about the potential for growth models to provide an alternative or useful addition to the accountability systems that each state is implementing under the requirements of the No Child Left Behind (NCLB) act. Policymakers and educational leaders are also seeking more information as to the basic differences between various types of growth models and the essential assessments, data, and reporting systems that are needed to implement them. The purpose of this paper is to define the key differences between the statistical models that can be used for school accountability and to examine the various factors important in making decisions about incorporating a growth model into an accountability system. The present paper is one of several efforts by CCSSO to provide information to policymakers and education leaders interested in making informed decisions about growth models for accountability. The paper was commissioned by the Accountability Systems and Reporting (ASR) state collaborative and it was preparing by a writing team of consultants, state members, and staff.[1]

The ASR state collaborative was organized by CCSSO in 2000 to assist state education leaders to develop and improve state accountability. Members of the collaborative are staff of state education agencies charged with designing, planning, and implementing accountability systems. With the assistance of expert consultants and university-based researchers, CCSSO staff facilitates and coordinates the activities of the collaborative. As part of its mission, the collaborative has produced a series of papers that advise state leaders on key decisions for state accountability including critical issues with AYP under NCLB, improving state reporting and report cards, the validity of accountability systems, and a review of state amendments to NCLB/AYP (see References). Each paper combines an analysis of existing research with examples of best practice from the field to provide recommendations to state leaders within the larger context of a nationwide perspective.

Based on requests from member states, the ASR Collaborative engaged in this analysis of growth models by building on several activities of CCSSO to improve available information and to help guide discussions of growth models, including the following:

- a November 2004 meeting of state education leaders and researchers that provided information on existing accountability systems that include a growth model or component (for papers and presentations, see www.ccsso.org/projects/Accountability_Systems/Resources/#growth)

- a January 2005 meeting of leaders from states, national organizations, and the U.S. Department of Education (ED) to identify the key issues and questions about growth models (see same webpage)

- participation of chief state school officers and state education leaders in the U.S. secretary of education's special task force on growth models in accountability

This paper addresses many questions education leaders may have about the differences between status models and growth models. Both status models and growth models used for school accountability are defined and described. Additionally, a type of status model (the improvement model) and a type of growth model (the value-added model) are discussed in the paper. The paper then goes on to provide more specific information about current research and practices regarding the different models:

- purposes and structures of status and growth models for accountability, with the focus on school as the unit of analysis

---

[1] ASR and CCSSO appreciate the excellent work and collaboration of the authors: Pete Goldschmidt and Kilchan Choi (consultants), UCLA/CRESST; Pat Roschewski, Nebraska Department of Education; William Auty (consultant), EdMeasure; Steve Hebbler, Mississippi Department of Education; Rolf Blank and Andra Williams, CCSSO.

- advantages and disadvantages of the different models in relation to purposes of an accountability system
- challenges in implementation of growth models and value-added models (VAMs) and description of resource requirements
- policy questions that may be addressed prior to state policymakers choosing to use a growth model for education accountability
- considerations in deciding to combine a growth model with a status model, such as the AYP reporting requirements under NCLB

This paper addresses the potential use of growth models for school accountability and the possibility of adding a growth model to existing systems to provide additional information about educational performance of schools and groups of students. For example, a growth model can have the purpose of predicting whether and when a school will meet a projected proficiency goal. The paper does not address other common uses of growth analysis models, such as for diagnosis of individual students, for teacher evaluation, or for evaluating longitudinal effects of specific programs.

Section I of the paper defines the different accountability models; Section II addresses typical policy-related questions concerning growth measures; and Section III describes research behind status, growth, and VAMs, provides examples of where and how the models have been applied, and displays the various models using a matrix to compare specific characteristics of the models and the differences in data system and resource requirements. A glossary of terms and definitions is also provided.

I: Definitions of School Accountability Models

Status models are often contrasted with growth models. A status model (such as Adequate Yearly Progress [AYP] under NCLB) takes a snapshot of a subgroup's or school's level of student proficiency at one point in time (or an average of two or more points in time) and often compares that proficiency level with an established target. In AYP, that target is the annual measurable objective (AMO—the level of proficiency the state established as an annual goal for schools and students). Therefore, progress is defined by the percentage of students achieving at the proficient level for that particular year, and the school is evaluated based on whether the student group met or did not meet the goal.

Figure 1: Status Model
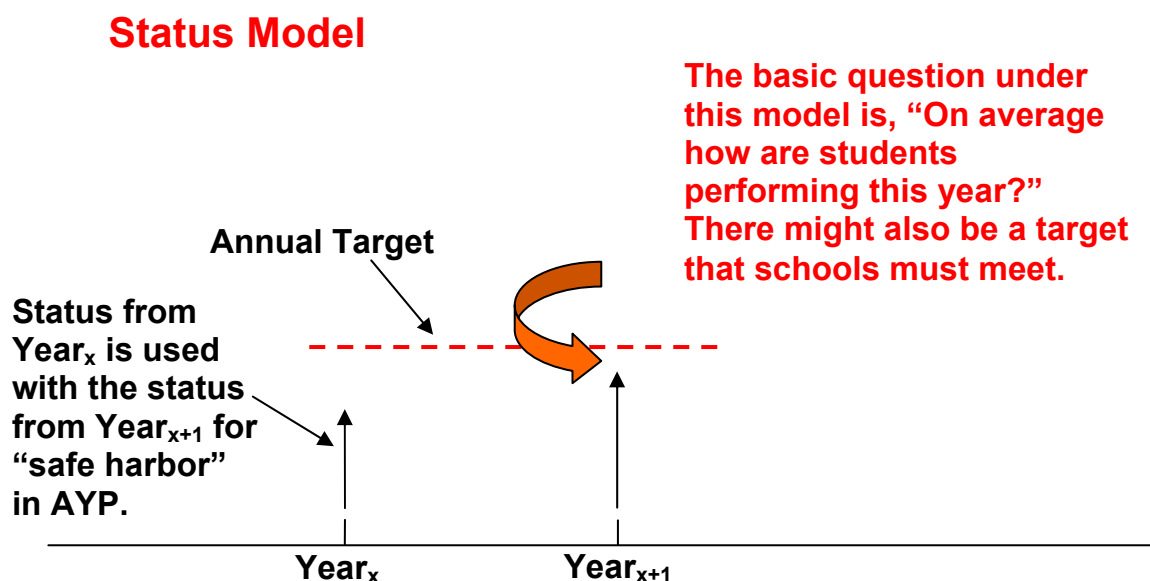
A status model analyzes school educational achievement compared against an established performance target—usually for one specific school year. In addition, status can be compared at two points in time to provide a measure of improvement. An **improvement model** of accountability is a type of status model which measures change between different groups of students (e.g., the performance of this year's fourth graders compared with last year's fourth graders). Such tracking of changes in proficiency levels is used as part of the AYP designations within the "safe harbor" provision of NCLB (which applies when the number of below proficient scores of a student group decreases by 10 percent from the prior year's comparable student group).

Figure 2: Improvement Model

Improvement Model

Improvement

The basic question under this model is, "On average, are students doing better this year as compared to students in the same grade last year?"

Year$_x$ Year$_{x+1}$
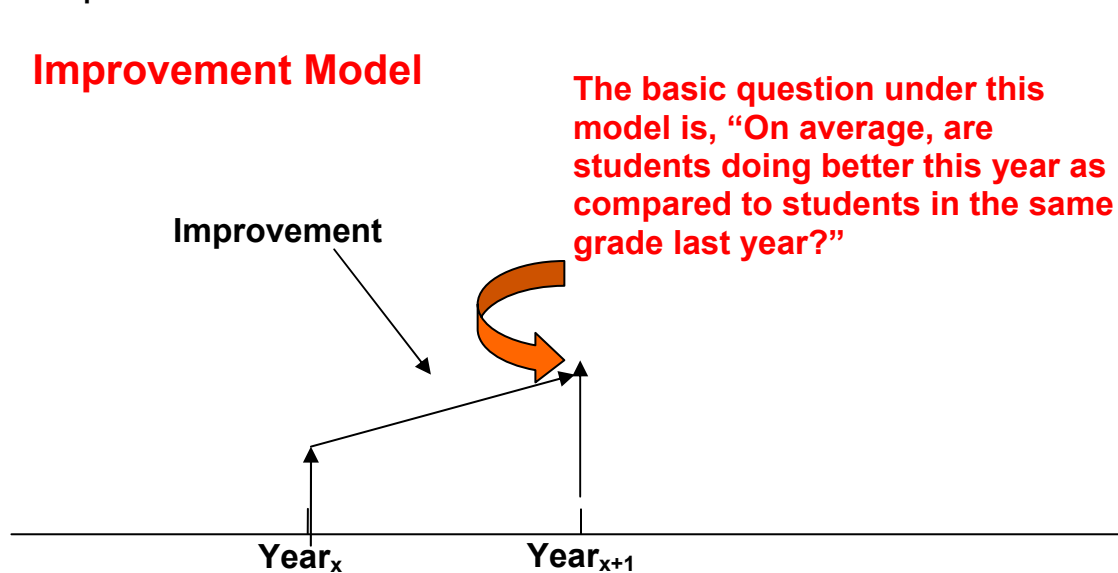
Growth models generally refer to models of education accountability that measure progress by tracking the achievement scores of a the same students from one year to the next with the intent of determining whether or not, on average, the students made progress. For example, learning growth can be measured by comparing the performance of this year's fourth graders with the performance of the same students last year in third grade. Achievement growth over time at the school level is then the aggregate of growth for individual students, controlling for each student's background and prior achievement. By comparing data for the same students over time, progress can be defined as the degree to which students' estimated improvement compares to a statewide or local target.

Figure 3: Growth Model

Growth Model
(simplified "generic" example)

Performance after a specified period of time (e.g., 1 school year)

The basic question under this model is, "How much, on average, did students' performance change?" There might also be a target that schools must meet.

At least two scores for each student are necessary. A starting point (which may be more than one year earlier) is important in a growth model

Year$_x$ Year$_{x+1}$
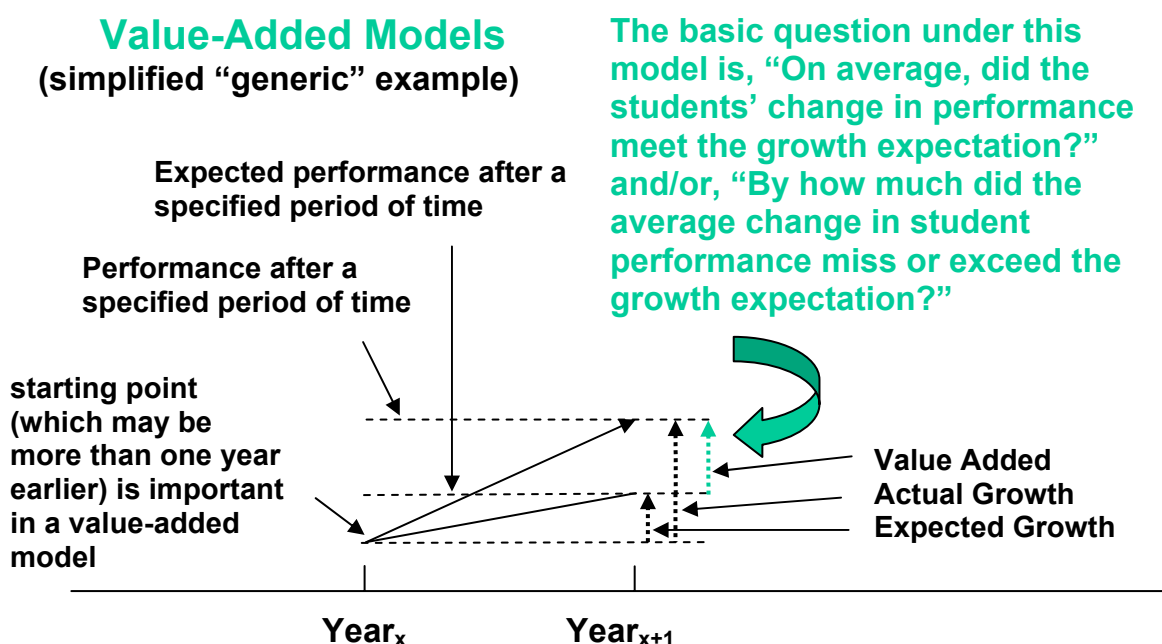
Growth models assume that student performance, and by extension school performance, is not simply a matter of where the school is at any single point in time, and a school's ability to facilitate academic progress is a better indicator of its performance. Growth models can vary, but in general, account for the potentially negative spurious relationship between status and growth, for status' effect on growth, and for student inputs' effect on growth. The greater the number of occasions (years) used to estimate growth, the less initial performance will be related to growth (Goldschmidt, 2004)—this means growth will be less and less related to indicators of school performance that are based on cross-sectional indicators (e.g., AYP). Schools can be ranked based on their growth estimates. In general, we would expect all students to demonstrate some academic progress across grades, but some schools will still exhibit more growth than others, on average.

A commonly referenced application of a growth model is a **value-added model**. VAMs are one type of growth model in which states or districts use student background characteristics and/or prior achievement and other data as statistical controls in order to isolate the specific effects of a particular school, program, or teacher on student academic progress[2]. The main purpose of VAMs is to separate the effects of non-school-related factors (such as family, peer, and individual influence) from a school's performance at any point in time so that student performance can be attributed appropriately. A value added estimate for a school is simply the difference between its actual growth and its expected growth. It is important to note that schools can demonstrate positive achievement growth, but still have a value-added estimate that is negative (i.e., the school demonstrated growth, just not as much as we would have predicted given the student inputs available to the school).

Figure 4: Value-Added Models

Value-Added Models
(simplified "generic" example)

The basic question under this model is, "On average, did the students' change in performance meet the growth expectation?" and/or, "By how much did the average change in student performance miss or exceed the growth expectation?"

Expected performance after a specified period of time

Performance after a specified period of time

starting point (which may be more than one year earlier) is important in a value-added model

Value Added
Actual Growth
Expected Growth

Year$_x$ Year$_{x+1}$

A well known type of value-added model is the Tennessee Value-Added Assessment System (TVAAS). Like most growth models, TVAAS tracks the yearly growth in student learning. However this model measures student growth by modeling a series of gains in performance demonstrated by each student as well as the teachers who instructed them and the schools that provided the context for their instruction. Thus, the model attempts to attribute the change in performance of students to the specific providers of

---

[2] We make the distinction between growth and value-added models by noting that when we use results from growth models, we focus on fixed effects estimates, and when we use results from value added models, we focus on random effects estimates.

instruction during a specific time period. While proponents of the VAMs view these links as opportunities for new levels of teacher accountability, there is little consensus on the issue. Although many scholars agree that VAMs can provide results from which to infer the effect of a classroom or a school, there is less agreement that TVAAS or other models can be used to accurately distinguish the effects of a single teacher.

II: Policy-Related Questions about Growth Measures in State Accountability Systems

Haven't we been measuring growth with our state tests?

Yes and No. Most states have had assessment systems in place that test student achievement at specific grades once each year. If the scores went up from one year to the next, those results were generally reported as growth. This measure of improvement is reported capturing differences in scores from one year to the next, but it does necessarily mean students actually learned more.

For example, if in 2002, 65 percent of third graders met the standard in mathematics and then in 2003, 70 percent of third graders met standard, that would reasonably be celebrated as evidence of growth in student achievement. However, the 2002 third graders are different students than the 2003 third graders, so the increase in scores may be attributable to more learning during the year, or maybe just the different characteristics of the different groups of students. Annual testing at a given grade is called a **status** model of assessment. One could say that this model provides a snapshot of achievement because we don't see the change in the same students' achievement as they progress from grade to grade. A **growth** model can measure the change in the proportion of student proficiency over time, the change in third graders' performance over time (a cohort model), or a change in individual student performance over time. The last model provides the most concise picture of what is happening to students as they progress through a school.

How could a state more accurately measure growth?

Individual student records are important. In the above example, we don't know how the 2002 third grades scored as fourth graders in 2003. To measure growth, the state would keep the 2002 scores of individual students and then give those students a test of the same content area in 2003. One would infer growth for an individual student if the score in 2003 is higher than the score she/he received in 2002. One would infer growth for a school or district if on average the test scores students received were higher in 2003 than what they received in 2002.

Is growth the best measure of student achievement?

Not necessarily: growth is different than status measures of achievement. This chart provides a simple picture that compares growth and status. Research in places that have tracked both growth and status reveals that schools can be found that represent all four categories. We would like all schools to produce both high growth and high status (Group IV) and we know that schools in Group I need to improve.

High Growth	High Growth
Group III	**Group IV**
Low Status	High Status
Low Growth	Low Growth
Group I	**Group II**
Low Status	High Status

However there are many schools that have produced mixed results and are in Groups II and III.

Measuring growth adds important information to evaluate the success of schools. Schools in Group II have scores that appear good. However students are entering the school with high achievement and leaving the school with achievement that is still acceptable, but there is little additional learning being produced during their time in the schools. Schools in Group III have not reached the desired levels of overall achievement, yet the students are making exceptional progress during their time in the school.

How should growth be measured?

There are many ways to measure growth. A basic type of growth model computes growth by subtracting each student's previous year's score from the student's current score. The result is the student's growth score. The growth scores of all students in a school can be averaged to get a growth score for the school.

Researchers have developed and tested different types of growth models that can be used for school accountability, and some are in use in states and districts (see References). A growth model provides information about students' score growth over time, and in addition, incorporates student or school

characteristics to provide a more accurate measure of the degree to which student score growth can be attributed to teaching and learning in the school. A growth model can be viewed as providing more accurate information about a school's success. One type of growth model is a value-added model. A value-added model isolates school measures of growth by taking into account factors such as student prior achievement, family background (e.g., ethnicity or income status), current class size, or teacher experience.

What is the best growth model?

The answer involves a policy decision. In general, the models that are designed to consider more factors (i.e., more precise models) are also more complex to compute and to explain. Therefore before choosing a growth model, it is important to be clear about why growth is being measured. A clear statement of policy intent allows a state to select a growth model that meets identified aims without being needlessly complex.

Is growth more expensive to test?

Probably, given that measuring growth requires more complex systems than using a status model. However whether a state will incur additional expenses depends on the infrastructure that it already has in place. For example, students need to be assessed with tests that produce comparable results from grade to grade and from year to year. (This is referred to as a vertical scale, or vertically equated tests, an important step in using a growth model for school accountability [Goldschmidt, Choi, and Martinez, 2003].) Prior to NCLB, some states had tests with different purposes and results at different grades. While NCLB does not prohibit that type of testing, it is no longer as practical or efficient under NCLB rules. On the other hand, states that are implementing new tests may have to wait a few years before they have enough data to assess growth from one year to the next.

Another possible expense is setting growth standards. As noted above, some schools will have high achievement and low growth while others will produce high growth without meeting the achievement status target. Therefore states will have to go through the standard setting process to establish AYP standards for both growth and status. If this process can be integrated into an already budgeted standard setting process, there may not be significant additional cost. However if the status standards are set, a state should estimate that setting growth standards might cost almost as much to establish as did the status standards.

A third cost factor is the availability of psychometric expertise. There are challenging technical issues to be resolved in creating a growth formula that meets a state's policy needs. If the state has adequately trained psychometric staff that can be assigned the task, then there may be little additional cost. If that resource is not available, then it must be obtained from out side. This could be a separate contract or an added expense to the state's existing assessment contract.

A fourth cost factor is the data system requirements. To measure growth, there must be a capacity to track individual student scores from one year to the next (and sometimes from one district to another in the state). This capacity often requires a statewide student identification system. About a dozen states have implemented statewide identification (ID) systems and many other states are in the process of developing them. There are many good reasons other than growth modeling to do so, but it can be expensive and time consuming. If the state already has or has budgeted for the development of an individual ID system, there will not be any additional cost to use those ID's for the growth model. However it is significantly more difficult to measure growth without a statewide ID, and if they are not in place, establishing a growth model will require allocating resources for a new ID system. So a state may have all the data system pieces in place to develop a growth model for little additional expense. However if any of the components are missing, the resources to obtain or develop them must be allocated.

A final cost factor is the training required to build capacity among the teachers, administrators, media, legislators, and general public to understand the additional complexities that occur when using data from more than one point in time. Even changing to the simplest of growth models will require a significant retooling of training materials. In addition, new validation reports may need to be created to allow schools to affirm that the correct matches have been made which may assist in increasing the model's credibility.

Does NCLB allow growth to be used to calculate AYP?

Not at this time. The current interpretation of the law by ED is that NCLB requires AYP to be based on the percent of students at a given grade meeting a predetermined target each year. While growth models are not specifically excluded by law, they are not currently allowed to mitigate the AYP rating. For example, an accountability system that includes growth could label as meeting AYP a school that missed the target if the students showed exceptional growth (Group III). Such a system could not be approved under current rules.

In the summer of 2005, ED began a series of meetings designed to produce guidelines for growth models that would meet NCLB requirements. It is anticipated that states will be able to submit accountability systems that compute AYP using growth for approval in time for the fall of 2006. However there are no plans to amend the law. Therefore some growth models currently being used or proposed are unlikely to be approved under the new guidelines.

Given the uncertainty and expense, why test growth?

Information is power. More information is more powerful. The achievement status of a school is not a sufficient indicator of success. Parents and other concerned citizens want to know if the school is improving the achievement of all students as they move through the school. A second reason often given by principals and teachers is that including growth in accountability is fairer than current systems. Local staffs are more willing to be held accountable for the growth a student makes as a result of instruction provided in the school rather than all of the things that did or did not happen to the student prior to that instruction (factors outside of school's control).

What are the key policy issues related to measuring growth?

There are many. States considering the addition of a component of growth in their accountability systems should analyze carefully the balance between the state's defined purposes for the growth model and the costs and other implications of doing so. Consider the following:

- *Acceptance under NCLB*: Accountability systems that include growth as a component of calculating AYP have not been approved by ED. New rules may allow some form of growth to be included in the future. States must make a policy choice whether or not to develop growth models with the sole purpose of measuring school performance that is valued in the state or to accept the constraints that come with federal approval under NCLB.

- *Setting growth standards*: There are implications for trying to set standards for accountability systems that include both growth and status components. This issue has not been approached in the education communities of states. Any standard setting process will have to address the conflicting values that will emerge. Group III schools may include some of the most prestigious schools in the state. These schools will have received positive ratings from existing status measures. Will it be politically acceptable to lower their ratings because of low growth? Group II will include schools that are producing exceptional results with traditionally hard to teach student populations. There will be difficult discussions about how much growth must be demonstrated to rate a school as acceptable when it hasn't met its achievement status target. The standard setting process will have to resolve honest differences in professional opinion and public values to include growth and status into a single rating system.

- *Factoring external performance standards into growth expectations*: Recall that growth models include a comparison of the increases in student performance against a specific target. A key policy question that must be addressed is how such targets are set. Expectations for growth can be based on past or typical growth (norm-referenced) or based on external performance goals (criterion-referenced). The factors that must be included in the decision are the balancing factors of capacity and sufficiency. For instance, a school principal might propose setting a growth target for a group of low-performing students that is equal to the average growth in the school (a norm-referenced target). The rationale would be that for many years, the low-performing students grew at less than half the rate of other students. So setting the target growth to be equal to other students would require doubling the growth of

the low performing students. While this proposal might be a reasonable goal in light of the perceived capacity of the school and its resource constraints, notice that the achievement gap would never be closed. To close the achievement gap, students who enter a school with low levels of achievement must grow at a greater than average rate. In contrast, a criterion-referenced growth target emphasizes sufficiency over capacity and establishes the amounts of growth needed to have all students reach a set achievement score by a set date. This is the method used under NCLB, and educators' concerns about the unreasonableness of those targets are well-documented. The state must resolve the policy question of how to set growth expectations that will produce the results the public wants while addressing the capacity of schools and educators to increase achievement.

- *Resolving technical issues*: Much of the rest of the paper provides detail about the emerging research and development in the area of growth modeling. Enough is known to provide policymakers with methods for improving accountability systems and thereby improve student achievement. However each state has unique circumstances (physical, social, technical, political, and historical) that must be considered as part of decision making process regarding the use of growth models in accountability. Therefore it will be important for states to establish good communication between policymakers and people with sufficient technical expertise that can make useful recommendations based on the latest research available. If this communication is effective, accountability systems are much more likely to produce desired results.

III: Key Technical and Practical Issues for Implementing Different Accountability Models

An accountability model is a systematic method of summarizing school performance. The model often becomes the basis upon which inferences about school performance are generated. Valid inferences depend on other elements beyond simply the certain accountability model that is used to generate results. For example, careful consideration to standards, assessments, and the alignment of assessments to standards needs to be undertaken before any model is applied. Failure to do so will lead to invalid inferences regarding school performance.

Section II outlined some of the key questions for the selection of an accountability model that must be considered. As noted, these are not necessarily based on either theory or empirical evidence; some are policy decisions based on more pragmatic issues like state or local political context and data system capacity. For example, the political climate may not allow for adjusting school performance by student enrollment characteristics. Or data systems may place physical constraints on the model choice. This section analyses the differences between growth models and seeks to answer several key questions that are often asked by policymakers and education leaders:

- Is a growth model preferred to a status model? If so, for what purpose?
- How are schools identified as needing improvement under a status model (e.g., AYP) vs. a growth model?
- Can a growth model be effectively combined with a status model within a state accountability system?
- What are the advantages of accountability scores that reflect results of school policy and practices? How does growth modeling improve the association between accountability and schools' policies?
- Should school accountability results be adjusted for differences in student characteristics?
- What type of model can better analyze effects of education in a school on its students?
- What type of model can identify performance gaps among student subgroups?
- What research supports use of growth models?
- What are the advantages of VAMs?
- Should school performance be judged against absolute or relative criteria?
- How do different VAMs compare?
- What are the data requirements for using growth or VAMs?

Currently NCLB is the common element of accountability models among states. While NCLB requires that states meet a target of 100 percent student proficiency in mathematics and reading by 2013–2014, many details are left to each state. Operationally, this means that while NCLB is the basis for all states' accountability models, states vary in their actual design and use of this model. For example, some states have simple linear trends, while others have stair-step patterns toward 100 percent student proficiency. Although a state's AYP model is based on progress toward 100 percent proficiency in 2013–2014, it is not a growth model because performance is evaluated yearly based on that year's performance.

How are schools identified as successful or needing improvement under a status model vs. a growth model?

School accountability based on meeting AMOs defined by the status model used in AYP measures for NCLB may not correctly classify school performance. This occurs for several reasons.

One reason is that as an artifact of construction, schools with more subgroups represented are more likely to miss meeting AYP due to the greater number of AMOs they need to meet (Novak and Fuller,

---

2003). A second reason for misclassification is that classifications based on a cut score capture only a small proportion of students' performance, especially when scores are close to the cut score (Thum, 2003). For example, a student whose scores are 1 point away from the target are treated exactly the same as a student that is 20 points away from the target. Similarly, AYP does not recognize that each student has an educational history and performs based on current and past opportunities to learn skills and build knowledge. A test from a student with many opportunities to learn is treated the same way as a test from a student without such advantages. Further, according to AYP, school performance is heavily influenced by the characteristics of the students who enroll in the school rather than how well the school instructs its students. For example, a school that happens to have 20 percent of its incoming students classified as gifted and talented will have better average performance than a school that has only 5 percent of its incoming students classified as such. Further, a model classifying schools based on a cut score will not provide a good indicator of school quality (Choi, Goldschmidt, and Yamashiro, in 2005).

Table 1 displays for easy comparison the key differences between status, growth, and value-added models of school accountability.

Table 1: Status, Growth, and Value-Added Models

General considerations by accountability model	Status models	Growth models	Value added models
Currently approved by ED for NCLB	Yes	No	No
Underlying purpose	Rank/rate schools based on current performance	Rank/rate schools based on performance change	Rank/rate schools based on performance changes different from expected
Major issues for consideration			
Results aligned with AYP	Very likely	Less likely	Less likely
Ratings generally understood	Very likely	Likely	Less likely
Inferences same as AYP	Yes	No	No
Requires more than 1 year of data	No	Yes	Yes
Unique student ID required	No	Generally	Generally
Potentially confounds student & school effects	Yes	Less likely	Less likely
Implementation time	Quick	Moderate/varies	High/varies
Implementation process	Simple	Moderate/varies	High/varies
Optimal testing requirements	None	Annual/same content	Annual/same content
Estimate teacher effects	No	Possible	Possible
Possible to measure within school inequities in performance	Limited	Possible	Possible
Costs	Low	Moderate	Potentially high
Simultaneously suitable for program evaluation	Unlikely	Yes	Yes
Measures change for individual students	No	Yes	Yes
-Absolute	-	Possible	Possible
-Relative to standard	-	Possible	Possible
-Requires equal interval scale	Yes	Yes	Yes
-Requires vertically equated scale score	No	Varies	Varies
Successful school profile	High average achievement, or exceeds % proficient target	High average achievement growth given average student enrollment	Higher than expected achievement growth given average student enrollment
Intended consequences	Reward high performing school	Rewards growth	Rewards better than expected growth
Unintended consequences	-Fosters status quo -Ignores within school inequities -Rewards schools with "favorable" enrollment -Does not reward student achievement growth (school improvement) -Reduce incentives for high quality teachers to teach	-May ignore high achieving schools -May ignore within school inequities -Perceived different standards for different sub-groups	-May ignore high achieving schools -May ignore within school inequities -Perceived different standards for different sub-groups

How does growth modeling improve analysis for accountability?

Simply comparing mean school performance over time (as in the AYP model) may indicate that student performance is very stable, but this does not ensure that performance is due solely to school policies and practices (Goldschmidt, 2004). As noted above, results based on the AYP model are confounded between school and student effects. In fact, the student characteristics that comprise the school may be contributing to the stability of the scores rather than the effects of school policy. Accountability based on models that simply aggregate performance will classify schools based on a combination of effects that are both within and outside of school control; however, school policymakers should be interested in isolating factors within schools' control for purposes of rewards, sanctions, and replicabililty (Willms and Raudenbush, 1989).

However if the characteristics of students in a school are included in an accountability model, school performance is not stable over time (Linn and Haug, 2002). If the state intends for the accountability system to classify schools based on factors under school control—that is, answering the question of how well an average student would perform in a given school—then adjustments are necessary (Rumberger and Palardy, 2004). Adjusting for student background can be accomplished in two ways. One is to explicitly incorporate students' backgrounds into the accountability model, while the other is to use a growth model rather than a status model and incorporate students' initial academic achievements into the model (Choi, Yamashiro, Seltzer, & Herman, 2004).

A growth model explicitly connects each student's performance from one year to a subsequent year. (Comparing average school performance from one year to the next, as in AYP, is not a growth model, rather a comparison of multiple statuses.) Growth models can either monitor cohorts as they pass through a grade or individual students as they move along grades. A status model has less data requirements than do growth models, but does not allow for clearly differentiating environmental factors from student factors. Following individual student scores is preferable over status models because growth models can incorporate each student's starting point that reflects their own unique history (Raudenbush and Bryk, 2002) and makes it possible to separate systematic schoolwide effects from individual student differences (Choi, Goldschmidt, and Yamashiro, in 2005). See Tables 2 and 3 for further details on Growth vs. Status Models.

Even under the most controlled conditions, determining cause and effect can be difficult. Unfortunately for researchers, schools are not controlled conditions and therefore are problematic when attempting to determine the specific cause of changes in student performance. Just because one policy change preceded a change in student performance does not mean that there was one and only one force acting on the students.

For example, based on school mean performance, a relationship between the percentage of students in a school eligible for free and reduced priced lunch (FRP) is inversely related to mean school performance. This could lead to the unsubstantiated conclusion that FRP students are not performing as well as their non-FRP eligible classmates. However, if in addition the number of FRP students was highly correlated with the number of migrant students and the number of migrant students was also inversely related to student performance, the observed relationship between FRP and student performance could be driven by the number of migrant students in the school (i.e., the relationship between FRP and student performance would be spurious). Aggregate data cannot be used to identify the specific relationships in this case because we do not know whether FRP or non-FRP students are the actual low performers in low-performing schools. Using individual student assessment results and then aggregating these to the school is not an adequate procedure because it overestimates precision and reliability (Aitkin and Longford, 1986)

What are key differences within the status models category?

Status models can be categorized into unconditional and conditional status models. An unconditional model uses unadjusted mean school performance, or percentage of proficient as an indicator of performance. A conditional status model attempts to account for factors affecting performance that lie outside of a school's control. Generally, parents would be interested in results based on an unconditional model and policymakers would be more interested in results based on a conditional model (Willms and Raudenbush, 1989).

A status model assumes that irrespective of everything else that could possibly affect student academic performance, both currently or historically, the snapshot of current performance accurately reflects how the school is performing. That is, in a status model all student success is attributable to the current school in the current year. It assumes that students do not bring any "human capital" inputs with them to the school. For example, a more print-rich home environment has no bearing on student readiness for reading in first grade. By using an unconditional status model, one assumes that there are no selection effects; the students in school A are like any other students in any other school in the district/state. One could replace students in school A with other students from any other school and the indicator of school performance would not change.

A conditional status model also assumes that all student success in the current year is attributable to the current school in the current year. The major difference between a conditional model and an unconditional model is that the conditional model recognizes that students do bring "human capital" inputs with them into school. These inputs cannot be measured exactly but we often have reasonable proxies in the form of student background characteristics that are included in the conditional status model. By generating an indicator based on a conditional model we assume that the school's performance does not depend on student inputs and that replacing students from one school with students from another would not change the school's indicator of performance.

It is incorrect to assume that because conditional means are used in the accountability model that students and or schools will then implicitly be held to different standards. Individual student performance standards are now set independent of test results (i.e., a student must be proficient in order to meet NCLB requirements) and this standard does not depend on past performance or demographic background. The impetus for conditioning performance on student inputs is to be able to compare performance among schools with similar characteristics rather than provide an adjustment (excuse) for having more or less favorable student input characteristics. It is incumbent that states and district present conditional results in this manner (e.g., California's similar school index). Policymakers may be concerned that schools with less favorable student input characteristics will depend on the adjustment mechanism in the model to hide poor school performance, but purpose of this conditioning is to allow policymakers to compare schools with similar student input characteristics, thereby highlighting performance deficiencies rather than hiding them. For example, school A may claim that it demonstrated poor performance because it has a high proportion of students with disabilities (SWD), but if this is explicitly taken into account in the model, then schools with similarly high proportions of SWD can be compared with one another. This allows comparisons of like with like (Choi, Goldschmidt, and Yamashiro , 2005; Goldstein & Spiegelhalter, 1996; Meyer, 1996)

Status models that compare two student cohorts can provide legitimate information regarding school performance, as long as policymakers are aware that this type of model focuses on changes in academic performance of different cohorts. If data from multiple cohorts are available, it is possible to conduct longitudinal analyses that can also be used to monitor school performance over time (Choi, Goldschmidt, & Martinez, 2004).

What are the key ingredients of growth models?

Growth models are a subset of the more general longitudinal models that examine how outcomes change as a function of time (Singer and Willet, 2003). Growth models assume the outcome of interest (achievement, attendance, retention) improves, and that scores or data collected over several points in time can be modeled longitudinally.

The simplest form of a growth model is a model that uses gains from one year to the next as the outcome. Growth models with three or more time points typically model the individual outcome as a function of time where the model generates a growth trajectory for each individual (Raudenbush and Bryk, 2002). These models are more flexible than traditional repeated-measures designs because data need not be balanced nor complete (Singer and Willet, 2003; Raudenbush and Bryk, 2002).

VAMs are a type of growth model designed to explicitly account for both the accumulation of effects over time and the confounding effects of student background in identifying school performance due to factors controllable by schools. In general, if School A can facilitate better individual student performance than

School B for a random average student, School A would be considered as providing more added value (value added) to that student's education than School B. As noted above, we cannot assign an average student to attend a school, but must take into account that at any given time a student's performance is a function of several factors, including family background, innate ability, peer influences, schooling, and luck (Hanusek, 1996). It is important to note that at any given time we are examining the accumulation of all of these factors from when the student begins school to the time of the current analysis. The assumption is that if this accumulation of factors influences performance at the time of the current assessment it must have also influenced performance at the time of the previous assessment. Hence if we focus on the change in performance from the time of the previous assessment to the current assessment, we minimize errors due to ignoring accumulated effects over time (Hanushek, 1986).

Whether the accountability model is based on a simple aggregation of student assessment results or on a longitudinal VAM, the criteria upon which to base performance can be either relative or absolute. Relative criteria have fewer data requirements (Goldschmidt, Choi, and Martinez, 2003) and are derived from the results of the models (e.g., schools can be said to have positive value added if their estimated performance is above the district mean). Absolute criteria, such as those generated by the AYP model (annual target for percent of students meeting/above proficient) are more difficult to validate. In general, the decision is political, and any estimate as to sufficient value added, for example, depends entirely on the criteria that are set. We also note that schools can be judged against either relative or absolute criteria and that either standard applies to status models (e.g., AYP) or growth models (e.g., VAMs). It is also important to note that relative criteria will not necessarily create standards that are consistent with the expectations of stakeholders; absolute criteria provide the opportunity to build consensus around achievement expectations.

The concept that is not intuitive, as described in (Choi, Goldschmidt, Yamashiro, 2005), is that school can demonstrate gains, but not necessarily provide positive value added to the students. That is, based on its enrollment or the initial academic status of its students, a school would be expected to demonstrate a certain amount of gain. If a school exceeds this expectation, it is defined to have positive value added. The expected gain or criteria against which school performance is compared can be set a priori, or derived empirically from the data.

How do VAMs compare?

The differences in inferences based on different VAMs will be much less than the differences in inferences between a VAM and a status model such as AYP (Choi, Goldschmidt, and Yamashiro, in press). Status and growth are only marginally correlated, and becomes less correlated the more time points are used to estimate growth (Goldschmidt, 2004) making results based on these models less similar. Still, there are some differences among some of the currently more widely known VAMs. Based on (Choi, Goldschmidt, and Yamshiro, in press) we briefly summarize these differences in Table 4.

The Tennessee Value-Added Assessment System (TVAAS) (Sanders, Saxton, & Horn, 1997) incorporates multiple content measures as well multiple cohorts and panel data. It does not assume simple linear growth, and it implicitly adjusts for prior achievement by using the gain scores and layers prior years' achievement onto each subsequent year's achievement. This model assumes that the effect of prior years' achievement remains stable over time. In other words, teacher effects are assumed to be constant over time. The TVAAS model does not adjust for student background characteristics, though an extension using a two-stage procedure to include student and school characteristics has been proposed (Ballou, Sanders, and Wright 2004). Some of the challenges in the TVAAS are that it does not attempt to model the interaction between where a school starts and how much it grows. There are tremendous data requirements (linked student data over time) as well as tremendous computing capacity requirements. Further, due to proprietary estimation procedures, broad applications of this model independently by states are not possible. Hence cost is an additional factor. Further, using models that contain complex (and proprietary) computations which are inaccessible to stakeholders may make it harder to build consensus and a sense of confidence around the validity of the results.

A more general form of Sanders's TVAAS model is McCaffrey et al's RAND Model; i.e., a multivariate, longitudinal mixed model that incorporates the complex nested structure of student longitudinal data linked to teachers (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). This model can include adjustments for student and school characteristics. Unlike the TVAAS model, it does not assume that

prior teacher effects remain constant over time. This allows the model to test whether prior teacher effects differ from current teacher effects and the extent to which teacher effects fade over time. Some of the shortcomings with this model (and Sanders's TVAAS model) are that ranking teachers is likely to be difficult, because obtaining precise estimates of teacher effects requires a large data set (Ballou et al, 2004; also see Lockwood, Louis, & MaCaffrey, 2002). The RAND study found that both McCaffrey's and Sanders's model identified only one-fourth to one-third of the teachers as distinct from the mean. This problem is, in part, due to the fact that when modeling teacher effects, rather than school effects, small classroom sizes make precise estimation difficult. This implies that making distinctions between teachers based on their value-added estimates could be challenging. Like Sanders's TVAAS model, this model does not address the interaction of where a school starts and how much it grows. Wider application of this model may be limited by the complex estimation procedure used, model convergence problems, available software, and extensive computing time. In addition, states need to consider the potential ramifications of high stakes accountability systems that explicitly attempt to isolate teacher effects. Regardless of whether it is consistent with the intent of the policy, teachers in the field may react negatively to the perceived increase in pressure. Care should be taken to minimize the potential unintended consequences of such a change.

A slightly different approach from the two previous VAMs is the Chicago Public School Productivity (CPSP) Model. This model uses a "productivity profile" in which initial status trends (input trends) and gain trends (learning gain trends), as well as output trends (adding input and learning gain together) are estimated for each grade level (Byrk, Thum, Easton, & Luppescu, 1998). Productivity is judged by both the learning gain trend and the output trend, to address situations where gains are up but inputs and outputs are down (gap is spreading). One of the key strengths of the Chicago model is that it is designed around a well-designed testing system. This model estimates both initial status as well as the gain, and can be applied to multiple subject areas and multiple cohort data. The model also explicitly takes measurement error into account. It is also possible to adjust for student and school-level characteristics. The model can be fitted using an accessible and well-known software program (HLM6, 2004). However this model does not specify modeling gain trends as a function of initial status trends. Also, this model requires vertically-equated test scores. Further, there is some expertise and training needed to correctly setup data files and complete the analysis using HLM. Although the software is relatively inexpensive, there are potential long-term staffing requirements that should be considered as well.

Another approach is the CRESST model that uses a latent variable regression in a hierarchical modeling framework (Choi, Seltzer, Herman, & Yamashiro, 2004; See also Seltzer et al., 2003, and Choi & Seltzer, 2004). One of the strengths of this model is that it estimates average school growth as well as the distribution of student growth within a school by explicitly modeling student growth as a function of a student's initial status. Modeling the relationship between where a student starts and how much he or she grows provides complementary information about how equitably student growth is distributed within a school for particular performance subgroups. Like the CPSP model, the CRESST model explicitly takes measurement error into account when standard errors of measurement are available. As with the other VAMs, the CRESST model can be extended to incorporate multiple measures, multiple cohorts; and to adjust for observed student and school-level covariates. Computation limitations with very large data sets as well as the model's complexity might be a shortcoming of this model.

Another approach is to use value added results as a basis for *measuring* progress towards a goal (Thum, 2003a; Doran and Izumi, 2004). Still another approach is to use a hybrid model (Goldschmidt, Choi, 2005). It is important to note that models continue to evolve precisely because no single model can address every issue. For example, newer models simultaneously estimate growth attributable to individual students and to sequential cohorts, thereby allowing portioning school improvement into what schools do for the same students over time and subsequent cohorts over time (Hara and Goldschmidt, 2005). Table 4 summarizes the characteristics of these models.

Conclusion

If states want to carefully address whether the correct schools are being identified for rewards or sanctions, want to validate AYP results, or begin to address whether accountability results merely reflect aggregated student performance or sound school policy and practices, they will need to make several key decisions.

As noted at the outset, accountability model results are only as good as the policies and data that drive the system. If the assessments are not aligned to content standards, are not reliable, and do not produce validly interpretable results, no amount of manipulation can create sound indicators of school performance. States need to decide whether to use a status model or a growth model or some combination thereof. This decision is independent of whether the model should adjust for student background or not, and of whether performance should be judged against absolute or relative criteria.

Research has demonstrated that unadjusted status models are imprecise indicators of true school performance. To move beyond this type of an accountability model, states will need to gather and maintain additional data. If states want to make statements about student achievement growth, assessments will need to be vertically equated. If relative performance (comparing or ranking schools) is sufficient, an equal interval metric will suffice. If student adjustments are desired, then these data must be collected and maintained (i.e., language proficiency status changes over time). To that end a universal, unchanging, student ID is needed, and the ID must follow students if they change schools. Assessments should be given annually in the same content areas. Models can adjust for changing assessments, but changing assessments annually confounds student growth with test effects.

Finally, states need to not only have computing capacity, but also the human capital to manage and understand the accountability model's results. Data systems developed to run multilevel longitudinal VAM models can generate more than simply a single value-added estimate for a school. In order to benefit from ancillary results, states must have personnel to examine results carefully. The alternative is to purchase a ready-made system that may be costly and create a "black box" atmosphere of distrust. Furthermore, at some point a ready-made model may not serve the state's particular needs and the state may be limited in its revision options. In any case, growth and VAMs can potentially provide an array of information, but should not be regarded as the only indicators of school performance.

Table 2: Growth Model Characteristics by Performance Indicator

Characteristic of Model	School performance indicator Growth Model		
	Gain score	**Conditional gain score**	**Multiple time-point growth**
Underlying purpose	Potentially confounded by IS or SB	Conditional gain score	Conditional gain score
States currently using	Various	Various	No
Major tradeoffs			
Results aligned with AYP	Less likely	Less likely	Less likely
Ratings generally understood	Likely	Less likely	Less likely
Inferences same as AYP	No	No	No
Potentially confounds student & school effects	Likely	Less likely	Less likely
Implementation time	Moderate	Moderate	Moderate
Implementation process	Moderate, possible in-house	Moderate, possible in-house	Moderate, maybe in-house
Estimate teacher effects	No	Possible	Possible
Possible to measure within school inequities	Limited	Possible	Possible
Costs (implicit in-house or outsource)	Moderate	Moderate	Moderate
Simultaneously suitable for program evaluation	No	Yes	Yes
Successful student profile (Individual value added estimates inappropriate)	High achievement Growth	High achievement growth given average student attr.	High achievement growth given average student attr.
Successful school profile	High average achievement. Growth	High achievement growth given average student attr.	
Type of growth examined	Year 2 - Year 1	Year 2 - Year 1	Growth trajectory
Methodological advantages	Simplest measure of true gain	Estimate true gain, account for student background	Estimate individual growth trajectory, account for student background
Methodological disadvantages	Potentially confounding	Potentially confounding	Potentially confounding
Type of software program	Basic statistical package	Packages allowing random effects (e.g., HLM, SAS, others)	Packages allowing random effects (e.g., HLM, SAS, others)

Table 3: Status Model Characteristics by Performance Indicator

	School performance indicator Status Model			
Characteristic of Model	**Percent proficient (AYP)**	**Mean percentile**	**Mean NCE or scale score**	**Conditional mean NCE or scale score**
Underlying purpose	Tracks change around score	Not appropriate for aggregate or growth	Confounded by time & external factors	Confounded by time
States currently using	All	Various	Various	Various
Major Tradeoffs				
Results aligned with AYP	Yes	Very likely	Very likely	Less likely
Ratings generally understood	Very likely	Very likely	Very likely	Less likely
Inferences same as AYP	Yes	Very similar	Very similar	No
Potentially confounds student & school effects	Yes	Yes	Yes	Less
Implementation time	Quick	Quick	Quick	Moderate
Implementation process	Simple	Simple	Simple	Moderate
Estimate teacher effects	No	No	No	No
Teacher effect models in use	n/a	n/a	n/a	n/a
Possible to measure within school inequities	Limited	Limited	Limited	Possible
Costs (implicit in-house or outsource)	Low	Low	Low	Moderate
Simultaneously suitable for program evaluation	No	No	No	Possible
Successful student profile	High achievement	High achievement	High achievement	High achievement given average student attr.
Individual value added estimates inappropriate	High achievement	High achievement	High achievement	High achievement given average student attr.
Successful school profile	High average achievement	High average achievement	High average achievement	High average achievement given average student enrollment
Type of growth examined	None	None	None	None
Methodological advantages	Mandated	Simple	Simple	Simplest way to account for student background
Methodological disadvantages	Potential confounding	Potential confounding	Potential confounding	Potential confounding
Software	Many	Many	Many	Many

Table 4: Characteristics of Value-Added Models

Characteristic of Model	Type of Value-Added Model					
	TVASS Model	RAND Model	Chicago Public Schools Productivity Model	CRESST Gain Model	Progress Towards Goal	Hybrid VAM
Underlying purpose	Implicitly accounts for initial status	Implicitly accounts for initial status	Uses input & output trends	Explicitly accounts for initial status	Estimates current growth & estimated growth needed	Estimates current growth & estimates probability of pass
Current locations	Tennessee	None	Chicago, IL	Seattle, WA	None	None
Major tradeoffs						
Results aligned with AYP	Less likely	Less likely	Less likely	Less likely	Less likely	Less likely
Ratings generally understood	Less likely	Less likely	Less likely	Less likely	Less likely	Less likely
Inference same as AYP	No	No	No	No	No	No
Potentially confounds student & school effects	Less likely	Less likely	Less likely	Less likely	Less likely	Less likely
Implementation time	High/varies	High/varies	High/varies	High/varies	High/varies	High/varies
Implementation process	Difficult/In-house not possible	Difficult/In-house not possible	Difficult/In-house not possible	Difficult/In-house implementation possible	Difficult/In-house implementation possible	Difficult/In-house implementation possible
Estimate teacher effects	Yes	Yes	Yes	Possible	Possible	Possible
Teacher effect models in use	Yes	Estimated	Estimated	No	No	No
Possible to measure within school inequities	Limited	Limited	Limited	Yes	Limited	Possible
Costs	High	Unknown	Unknown	Unknown	Moderate	Moderate
Simultaneously suitable for program evaluation	Possible, not practical	Possible, not practical	Possible, not practical	Possible, not practical	Possible	Possible
Successful student profile	High output trends compared to input trends	High output trends compared to input trends	High output trends compared to input trends	Greater gains than expected given initial status	Est. growth	High probability of passing given
Individual value added estimates inappropriate	Higher than expected gain compared to norms	Higher than expected gain compared to norms	High output trends compared to input trends	Greater gains than expected given initial status	Est. growth	High probability of passing given
Successful school profile	Higher than expected average gain compared to norms	Higher than expected average gain compared to norms	High school output trends compared to school input trends	Greater gains than expected given initial status/flag pattern	Est. growth	High probability of passing given
Type of growth examined	Layered gains	Layered gains	Output trends	Conditional Year 2 - Year 1	Progress towards criterion	Probability of meeting criterion use NRT and pass/fail outcome
Methodological advantages	Multiple years initial status implicit	Teacher effects not additive	Well established	Estimate within school variation	Links growth to end	Use NRT and pass/fail outcome
Methodological disadvantages	Extreme complex - convergence	Extreme complex - convergence	Results with cross-classification	Complex modeling	No error in growth	Unknown
Software	Proprietary	Proprietary	HLM, other	WinBUGS	R, SAS	WinBUGS

Glossary

Conditional score	Scores that are conditional are scores that are adjusted based on a statistical model that accounts for factors of interest.
Confounded results	Results may be confounded by time or external factors (e.g., family, teacher, school, and community input factors) that are not explicitly measured or modeled. Potential confounding factors (PCF) represent rival hypotheses to those under study. For example, the research hypothesis may be that school A does a better job of facilitating student achievement growth than school B— PCFs present alternative explanations to the research hypothesis. Some argue that it does not matter why a school's mean scores are high (or low), but simply that they are. Others argue that a school should only be held accountable for that part of student performance based on inputs the school can control. Parents and educators may differ on this issue. Resolving the purpose for measuring or modeling growth will help determine which factors should be explicitly included in the model.
CPS model	The Chicago Public Schools (CPS) model is a simple gain model using two time points, and the model avoids a problem related to degree of freedom issue with two-time points by using standard errors of measurement. Each student's learning gain data in successive years was treated as independent, and the full longitudinal nature of the individual data was not directly used. In other words, it is repeatedly fitting three-level gain model for the consecutive years (e.g., between years 1 and 2, and years 2 and 3, and so on). The input trend is referred to the trend of status of starting years of the pairs of years and the output trend means the trend of the status of ending years of the pairs of years.
Equal-interval scale	An equal interval scale is a scale in which the differences between any two points along that scale (e.g., 12-10 and 67-65) have the same meaning irrespective of where along the scale the difference is taken. The number of points along the scale does not determine whether a scale is an equal interval scale.
ES	End Status (ES) refers to a student's estimated true achievement at the last measurement occasion in the model.
Gain score	A gain score is the difference between student test scores at two time points (i.e. year2 - year1 score). The average of these differences can be used to describe the average gains for a group of students. This is the simplest form of a growth model. If standard errors of measurement (SEM) are available then gain scores could be modeled directly in the model. Using a computed gain in a model likely provides the best estimate of true gain, but ignores both measurement error and the correlation of errors over time within persons. An alternative to gain scores (especially if the SEM is unavailable) is to use a covariance adjustment model that uses year1 score as a predictor in year2 score.
Growth trajectory	The estimated growth per unit of time over the period of time studied. For example if there are test scores for the same individual student for grades 2,3, and 4 and that student is estimated to demonstrate achievement growth of 20 points per occasion, then the growth trajectory for that student is 20 points per grade of the specified three year span.
Hybrid value-added	A Hybrid Value Added Model (Goldschmidt and Choi 2005) is similar to a Progress Towards Goal model except that the probability of passing – being at or above a specific cut score is modeled using estimated true growth (latent

	growth) and thus takes the variability of growth into account. This model also combines growth based on a continuous measure and a final outcome based on a dichotomous pass/fail (dropout/not dropout etc).
Improvement	An improvement model is a type of status model that compares student performance from one year with the same school or grade performance in a prior or succeeding year. These models are often referred to as longitudinal cohort, or school productivity models.
IS	Initial Status (IS) refers to a student's estimated true achievement at the first measurement occasion. Adjusting growth estimates for initial status explicitly takes into account where a student starts. It is preferable to use latent IS rather than simply using the first test scores as a covariate because the covariate is measured with error, which would be assumed to be zero, and uncorrelated with anything else in the model (which is unlikely); the latent approach takes these problems into account.
Multiple time-point growth	A multiple time point growth model uses 3 or more scores to measure growth. Growth is defined as the average change in (this case) test scores per time. Time can be measured in many ways, though most commonly, grade, year, or test-occasion, are used. These models can account for the lack of independence among errors. Growth models can also account for a student's initial status (see below) and its indirect effect on growth.
NCE	The Normal Curve Equivalent (NCE) was designed to be used in the evaluation of compensatory education and other special programs in which scores from different tests are to be reported together. NCEs range from 1 to 99 and are equal to the National Percentile scale at 1, 50, and 99. NCEs are normalized equal interval scores and are not recommended for use in reporting individual student scores since the NCE is easily confused with the NP.
NP	The National Percentile (NP) represents the percentage of students in a national norm group whose scores fall below a given student's score. For example, a student whose NP is 65 scored higher than 65 percent of the students in the norm group. NPs are scores that are useful for comparing local student achievement to students' achievement nationally. NPs are not equal interval scores cannot be aggregated to classrooms or schools, and cannot be used for longitudinal analyses.
Progress towards goal model	Progress Towards Goal models intend to determine how much progress towards a set goal or cut score a student has made over some period. These models estimate growth rates and then the probability of that rate of growth meeting or exceeding the designated cut score (Thum, 2003, Izumi and Doran, 2004). These models use multiple assessments on a vertical scale upon which to base results. Also, in some cases (Izumi and Doran, 2004) the probability (or ratio) uses as its numerator the difference between the cut score and the calculated End Status (ES). This difference ignores variability in the desired growth.
RAND model	The RAND Model is the same as the TVASS when considering school effects. The major difference between the TVASS and the RAND model is in estimating teacher effects. The teacher effects remain constant and are cumulative in the TVASS model. In the RAND model teacher effects may be modeled as diminishing from year to year. In other words, a teacher may have a greater effect on student achievement in the year immediately after the instructional year than on achievement three year's later.
SB	Student Background (SB) are characteristics such as gender, race/ethnicity, socio-economic status, etc. Evidence suggests that using IS or SB to adjust

growth scores yield similar results (i.e., where a student starts is generally highly correlated with his/her background).

Scale score	A Scale Score is a transformation of a raw score (number of items answered correctly) into an equal-interval scale. A SAT score of 700 is an example of a scale score.
Status	Status models consider student performance in a single year. Status models can be conditional (take input factors such as student demographics or educational history into account) or unconditional (do not take inputs into account).
TVAAS model	The TVAAS (Tennessee Value-Added Assessment System) model is usually called a "layered model" because models for later years of teacher effects build upon the layers from the earlier year. In other words, The effect of each year's teacher is added. Therefore, the teacher effect is what remains of the year-to-year gain after removing the district mean gain and the contribution of factors idiosyncratic to the student.
Value-added	Value added is the difference between a school's actual growth and its expected growth. Schools may demonstrate positive growth over the specified period but may still have a negative value added estimate if this growth was less than expected.
Vertical scale	A vertical scale requires linking scores across contiguous grades. One way to do vgfthis would be to include both 2nd grade and 4th grade items on a 3rd grade Mathematics assessment as anchor items to link test scores across the grades.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, *4,* 158-233.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Choi, K., P. Goldschmidt, and K. Yamashiro (i2005) Exploring Models of School Performance: from Theory to Practice in National Society for the Study of Education v. 104, Joan Herman and Ed Haertel Eds. Blackwell pub.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2004). Children Left Behind in AYP and Non-AYP Schools: Using Student Progress and the Distribution of Student Gains to Validate AYP. CSE Technical Report. Los Angeles, CA: Center for the Research on Evaluation, Standards, and Student Testing, University of California Los Angeles.

Choi, Kilchan, Goldshmidt, P., & Martinez, F. (2004). *Monitoring school improvement over years using a hierarchical model under a multiple-cohorts design: comparing scale score to NCE results.* Paper presentation, annual meeting of the American Educational Research Association, San Diego.

Discussion of Key Issues for States with Growth Models for School Accountability: Issues Include Equity, Inclusion, Legal, and Political. Panel presentation, November, 2004. Ross Wiener, Education Trust, available http://www.ccsso.org/content/pdfs/Weiner.ppt. Mark Davison, University of Minnesota, available http://www.ccsso.org/content/pdfs/Davison.ppt.

Doran, H., & Izumi, L. (2004). Putting Education to the Test: A Value-added Model for California. San Francisco: Pacific Research Institute.

Examples of Two State Models Currently in Operation. Panel presentation, November 2004. Christopher Cobitz, North Carolina Department of Education, available http://www.ccsso.org/content/pdfs/Cobitz.ppt. Mary Reel and Benjamin Brown, Tennessee Department of Education, available http://www.ccsso.org/content/pdfs/Brown_Reel.ppt.

Goldschmidt, P. K. Choi (2005). A Latent growth curve approach to predicting student proficiency. Paper presentation, annual meeting of the American Educational Research Association, Montreal.

Goldschmidt, P. K. Choi, F. Martinez (2003). Using Hierarchical Growth Models to Monitor School Performance Over Time: Comparing NCE to Scale Score Results, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), U.S. Department of Education, Office of Educational Research and Improvement

Goldschmidt, Pete (2004). Paper presentation, Symposium Large Scale Assessment Conference, Boston, MA. June.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A, 159*, 384-443.

Gong, B. (2004). Incorporating Student Growth (or Value-Added) Measurements Into School Accountability—What Are the Models? Center for Assessment. Available http://www.ccsso.org/content/pdfs/Gong.pdf.

Hanushek, E.,.A., Steven Rivkin, and Lori Taylor (1996).Aggregation and the estimated Effects of School, *The Review of Economics and Statistics*, v. 78(4), pp 611-27.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of education production functions. *Journal of Human Resources,* 14 (3): 351-388.

Hara, M & P. Goldschmidt (2005). Are there really good schools: the role of changing demographics within schools. Paper presentation, annual meeting of the American Educational Research Association, Montreal

Linn, R. L., and Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), 29-36.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D.W. Jorgenson (Eds.), Improving America's schools: The role of incentives (pp. 197-223). Washington, DC: National Academic Press.

Novak, J., & Fuller, B. (2003). Penalizing diverse schools? Similar test scores, but different students, bring federal sanction. PACE Policy brief. Berkeley, CA: Policy Analysis for California Education.

Other Examples of Use of Growth Models. Panel presentation, November 2004. Robert Bernstein, Rachel Perry, California Department of Education, available http://www.ccsso.org/content/pdfs/Bernstein_Perry.ppt. Mitchell Chester, Ohio Department of Education, available http://www.ccsso.org/content/pdfs/Chester.ppt. Steven Ponisciak, University of Chicago, available http://www.ccsso.org/content/pdfs/Ponisciak.ppt.

Pedhazur, E. (1982). Multiple regression in behavioral research. Fort Worth: Harcourt Brace Jovanovich

Rumberger, R. W., & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), Handbook of quantitative methodology for the social sciences (pp.235-258). Thousand Oaks, CA Sage Publications.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29* (1), 121-129.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed). Newbury Park, CA: Sage Press.

Raudenbush, S.W., Y.F. Cheong, and R.P. Fotiu (1994). Synthesizing Cross-National Classroom Effects Data: Alternative Models and Methods in Methodological Issues in Comparative Educational Studies, US Dept. Education, National Center for Education Statistics 94-469.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Implications for conducting analyses that help illuminate the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25* (3), 263-286.

Singer, J. and J. Willet (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press.

State Lessons for NCLB-AYP Year One: How Can Accountability Decisions Be Improved? (CCSSO conference on large scale assessment, 2004). Abigail Potts, Council of Chief State School Officers, available http://www.ccsso.org/content/pdfs/LSAC2004Potts.pdf. Mitchell Chester, Ohio Department of Education, available http://www.ccsso.org/content/pdfs/LSAC2004Chester.pdf. Juliane Dow, Massachusetts Department of Education, available http://www.ccsso.org/content/pdfs/LSAC2004Dow.pdf.

Thum, Y.M. (2003a). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research,* 32 (2), 153-207.

Thum, Y.M. (2003b). No Child Left Behind: Methodological Challenges & Recommendations for Measuring Adequate Yearly Progress. CSE Technical Report. Los Angeles, CA: Center for the Research on Evaluation, Standards, and Student Testing, University of California Los Angeles.

U.S. Public Law 107–110. 107th Congress. 2nd Session. 8th October, 2002. <u>The No Child Left Behind Act of 2001.</u>

<u>Use of Growth Models in School Accountability</u>, Memo to Chief State School Officers, March 2005. Available http://www.ccsso.org/content/pdfs/Growthmemo.pdf

Willms, D. & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability, *Journal of Educational Measurement*, 26(3), 209-232.