Implementer's Guide to Growth Models

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.

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

The State Collaborative on Assessment and Student Standards (SCASS) Project was initiated in 1991 to encourage and assist States in working collaboratively on assessment design and development of high quality assessments in relation to standards for student learning. State education agencies voluntarily support collaborative work across a variety of topics and subject areas. The Accountability Systems and Reporting (ASR) state collaborative project has been assisting States to develop and improve statewide systems since 2000. Currently 28 state education agencies are participating members, together with researchers, consultants, and measurement specialists. Priority activities of the ASR Collaborative are: state planning and development to meet requirements for state systems under NCLB, sharing strategies and research-based evidence to improve the validity of accountability systems, developing reporting formats that improve information access for constituents, and establishing standards for the essential components of accountability systems. Funding support for this paper was entirely from member States of the Accountability Systems and Reporting State Collaborative (ASR-SCASS). For further information about ASR SCASS and resources on accountability across the 50 States, see the CCSSO website: http://www.ccsso.org/projects/Accountabilty_Systems/.

2008
Council of Chief State School Officers
Rick Melmer (South Dakota), President
T. Kenneth James (Arkansas), President-Elect
Elizabeth Burmaster (Wisconsin) Past President
Gene Wilhoit, Executive Director

Rolf K. Blank, Director of Education Indicators

Council of Chief State School Officers
Attn: Publications
One Massachusetts Ave., NW, Suite 700
Washington, DC 20001
202-336-7016
Fax: 202-408-8072
www.ccsso.org

Copyright © 2008 by the Council of Chief State School Officers, Washington, DC.

All rights reserved.

Implementer's Guide to Growth Models

William Auty, Consultant

ASR Working Group
Paul Bielawski, Michigan
Tom Deeter, Iowa
Glenn Hirata, Hawaii
Christy Hovanetz-Lassila, Minnesota
Joanne Rheim, Delaware
Pete Goldschmidt, CRESST
Kimberly O'Malley, Pearson
Rolf Blank, CCSSO
Andra Williams, CCSSO

A paper commissioned by the CCSSO Accountability Systems and Reporting
State Collaborative Project

January 2008

Copyright © 2008 by the Council of Chief State School Officers, Washington, DC.
All rights reserved.

Table of Contents

Note to Readers of Electronic Versions of This Document:

Versions of this document in Word, pdf or html format contain hyperlinks to navigate through the text. The links are formatted like this: link and they provide quick ways to find the information you want.

Implementer's Guide to Growth Models

1. Descriptive Summary

There is currently great interest in using growth models to improve the quality of school accountability systems. Given that widespread interest, many state education agency (SEA) staff and policymakers are faced with making decisions about how to develop and implement a growth model in a statewide accountability system. Understanding and clarifying purposes and priorities of need for such a model can then guide decisions about the type of growth model and other implementation details. In this paper we describe the theoretical and practical issues that you will face in designing and implementing growth into your statewide accountability system. We also include examples of state growth models in use wherever possible.

The literature on growth models for accountability and school improvement is at an early stage of development and therefore the definitions of key terms have not been established. Through participation in ASR SCASS, state education specialists and consultants have shared their efforts to review and interpret the existing information on use of growth models in school accountability (see http://www.ccsso.org/projects/Accountability_Systems/), and much of the literature is from a few district and state applications of growth models. The first paper on growth models developed by ASR SCASS (*Policymakers' Guide to Growth)* reviewed and organized much of the existing literature (Goldschmidt et al, 2005). In this paper, Section 3 will define the various growth models as we did in the *Policymakers' Guide to Growth*. The reader should be aware that the terms might be used differently in other papers.

A key step is to understand the characteristics of each major type of accountability model so that reports of the implementation of growth models can be interpreted. 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. An **Improvement Model** of accountability is a type of status model which measures change between different groups of students. True **Growth Models** within educational accountability generally refer to models of education accountability that measure progress by tracking the achievement scores of the same students from one year to the next. In **Value-Added Models**, 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. The last defined model for growth is based on a **Transition Matrix**. In this model, growth is measured in relation to the performance categories, e.g. Basic, Proficient and Advanced. The advantage of this model is that it does not require a vertical scale.

Section 4 discusses the several possible purposes for using growth. States frequently run into problems when the reasons for implementing growth are not made clear. We can acknowledge that this step may not be perfectly completed, but it is still valuable to have policymakers create some record of their intentions. These purposes can then guide decisions about the type of growth model and other implementation details. Although there are many possible ways to implement growth models, there are a limited number of methods that underlie those models. Section 5 describes six methods and eight characteristics that differentiate the model types. Table 1 is provided to show how the characteristics vary across the growth models.

Although NCLB often refers to student growth, most accountability systems using true measures of growth could not be approved under the rules originally set by the U.S. Department of Education (ED). Now that the Department is exploring options for including growth, states may want to design their growth models to calculate AYP. The first design decision a state must make is whether to incorporate a growth component into its school accountability system. The second design decision is whether to use a growth model that meets the ED Growth Model Pilot's specifications. If the state decides it would like to be approved for the ED Growth Model Pilot, then the state must make design decisions about several other aspects, including the nine described in Section 6 of this document.

An important part of any assessment system is the need to effectively communicate the results. Because growth models are new and at times include complex statistical calculations, reports of student growth can be difficult to design. Section 7 describes eight guiding principles for high quality reporting. To show how states are reporting growth, in Section 8, Hawaii and Michigan describe how growth is to be reported in those two states. In addition, a project in Oregon shows us how to report on closing achievement gaps.

A question that quickly arises after deciding to measure growth is, "How much growth is enough?" This is a key question that must be answered before including growth in any accountability system. Most states have had experience with setting standards for individual student proficiency and those experiences can inform the process for setting standards for growth in an accountability system. However there are differences in procedures and in the participants in standards setting. Delaware and Florida provide us with two methods of setting standards for growth in their AYP systems.

Section 9 outlines the balance that implementers of growth models must find between policy goals and data availability in order to produce robust results. Robust results depend on resolving both technical issues such as precision, reliability, and stability as well as issues related to the validity of inferences. Growth models can provide substantially more information than status models, but can never the less benefit from considering both multiple analyses of the same data as well as multiple sources of information.

Many state accountability systems, including AYP, present to the public written, annual, cross-sectional results. For example, a report might include how third grade performance in a school changes from one year to the next. Given that both school improvement and multilevel models represent growth, a natural question might be the extent to which these models' results lead to the same inferences about school performance. This can be addressed using a system that simultaneously models improvement (changes in the subsequent performance of cohorts) and individual student growth.

Although there is a fortunate correspondence between recent developments in the measurement of growth and the desire for improved methods to hold schools accountable for the real difference they make in student achievement, there are unresolved technical issues related to using growth models. Section 10 includes technical notes that provide background information to help implementers resolve these issues.

We conclude with an annotated list of suggested readings in Section 11, which provide additional information on applying growth models to accountability systems. The literature on growth models is expanding rapidly. Readers are advised to look for new publications from organizations such as CCSSO, AERA and NCME.

2. Purpose of This Document

There are several possible purposes for developing and using a growth model within a statewide accountability system and for SEAs to develop and expand their capacity to analyze student-level data over time. States frequently run into problems when the reasons for implementing a growth model are not made clear. We acknowledge that the different purposes as well as the advantages, resource needs, costs, and benefits for moving into a statewide growth model vary by state and may not be clear at this time. However, it is still valuable to have policymakers create some record of their intentions. Understanding and clarifying purposes and priorities of need for such a model can then guide decisions about the type of growth model and other implementation details.

This document discusses the theoretical and practical issues that you will face in designing and implementing statewide growth models. We also include examples of growth models in use wherever possible. We also assume that few organizations will have staff with all the expertise needed to implement growth models. Therefore, information about when and how to contract for needed services is included. It is also important to note that no single growth model is best nor can any model address every situation.

3. Definitions of School Accountability Models

The literature on growth models is at an early stage of development and therefore the definitions of key terms have not been established. Through participation in ASR SCASS, state education specialists and consultants have shared their efforts to review and interpret the existing information on use of growth models in school accountability (see http://www.ccsso.org/projects/Accountability_Systems/), and much of the literature is from a few district and state applications of growth models. The first paper on growth models developed by ASR SCASS (*Policymakers' Guide to Growth)* reviewed and organized much of the existing literature (Goldschmidt, et al., 2005). For this paper, we will define the various growth models as we did in the *Policymakers' Guide to Growth*. The reader should be aware that the terms might be used differently in other papers. However, the key is to understand the characteristics of each model so that reports of the implementation of growth models can be interpreted.

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. This type of analysis is also referred to as **cross sectional** in that several layers or grades can be analyzed annually. An accountability model based on status monitors schools based on a single year's performance and generally rates schools either by ranking them or compared to an established target.

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. Status model results tend to be highly correlated with a school's

student enrollment demographic characteristics. A key issue with status models for many researchers is that they ignore the fact that learning is a cumulative process.

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). Improvement models can be used to estimate cohort growth (i.e. the change performance over time across cohorts). A benefit of an improvement model is that it has less stringent data requirements than growth models.

Growth Models generally refer to models of education accountability that measure progress by tracking the achievement scores of the same student from one year to the next with the intent of determining whether or not, on average, the student made progress. For example, growth in learning can be measured by comparing the performance of a fourth grader this year with the performance of the same student last year in the third grade. Achievement growth over time at the school level is then the aggregate of growth for individual students. 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. Growth models account for the cumulative process of learning by modeling achievement growth over time. An advantage of growth models is that students act as their own controls and thus account for student background characteristics. However, while some growth models may not require student input information (i.e. background or initial status), models that operate to provide projections will benefit from student input data. This is useful for accountability systems (e.g., NCLB) where such adjustments are prohibited or otherwise undesirable.

Growth models assume that a school's ability to facilitate academic progress is a better indicator of its performance than the NCLB status model of comparing the students' performance to an established target. 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. 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. The reason we use a growth model rather than simply subtracting scores from one year to the next is that, at the school level, growth estimates account for imprecision, missing data, and small sample sizes for subgroups.

An important element of growth models is that it accounts for the cumulative process of learning by modeling achievement growth over time. An advantage of growth models is that students act as their own controls and thus account for student background characteristics. This can be useful when there are policy concerns about setting growth targets based on past achievement of students with specific characteristics.

A commonly referenced application of a growth model is a **Value-Added Model**. In VAMs, 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. 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).

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 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.

Another model for growth is based on a **Transition Matrix**. In this model, growth is measured in relation to the performance categories, e.g. Basic, Proficient and Advanced. The advantage of this model is that it does not require a vertical scale. The assumption is that a student who scores in the proficient range at a given grade is making expected growth if he or she also scores proficient the following year. A value table can be constructed with the rows indicating performance categories for year one and the columns indicating performance categories in year two. The table cells indicate the possible changes in performance over the two years. The value associated with each can be entered into the cells. For example, one might give 100 points for maintaining the proficient level for two years and give 200 points for moving from basic to proficient. Typically, the points are determined by a standard setting process that captures the value of the accountability system's stakeholders. In Section 8, Delaware provides an example of using the transition matrix model as part of their AYP calculation.

4. Setting the Purpose for Using a Growth Model

There are several possible purposes for using growth and states frequently run into problems when the reasons for implementing growth are not made clear. We can acknowledge that this step may not be perfectly completed, but it is still valuable to have policymakers create some record of their intentions. These purposes can then guide decisions about the type of growth model and other implementation details. In this section we compare the broader purposes for growth models with some of the experiences with implementing models to this point.

A growth model can be used to examine the growth of every student along an achievement continuum, which could have important implications for modifying the AYP provisions of NCLB. For example, if the expectation is that all students, including those already proficient, make adequate growth, the questions arises, “What amount of growth is adequate yearly growth?” Should we expect the same amount of growth from each student, or should we expect more than a year’s growth in a year’s time if the student is below proficient? Conversely, is less growth in a year acceptable for students who have exceeded the proficiency standards? The current ED growth model pilot projects, much like the current status model, continue to examine growth for students who are not yet proficient, but who are “on track to be proficient.” The ED growth model pilot program did not have explicit

guidelines regarding growth for students above proficient. However, the last round did not allow states to run growth models for only those students below proficient (i.e. if a school did not make AYP, it could not run a growth model for only those students who were not proficient and then add those results to those for students who were proficient to get a proficient percentage. If a school used a growth model, all students would have to be included).

If the expectation is that all students should be monitored for growth, including students that have reached proficiency, the issue becomes more complex. We have to decide how much growth is enough for students at different levels of proficiency. Should we expect the same amount of growth from each student, a year's growth in a year's time, more than a year's growth in a year's time, or less growth in a year? Should all students improve at the same rate? Students who have scored high on tests in the past might actually not be expected to grow as much because of a "ceiling effect," in that students might have to work harder to achieve higher and higher levels of proficiency. Students who have scored at the low end of the distribution might be expected to grow more than others, because they have the most room to grow. In addition, the desire to close an achievement gap includes the expectation that low performing students would grow faster, i.e., "more than a year" in a year's time. However, they may also have to work as harder simply to master grade level concepts and maintain an average rate of growth. Therefore, we see that when the purpose for using a growth model is set, the decisions are complex and quickly lead to questions about the expectations for instruction and the support provided for classroom instruction.

Growth models *may* provide teachers and schools with evidence of achievement gains that they are producing, but don't show up in the status models. However, the jury is still out. Early returns from seven states implementing growth models under the NCLB pilot program (including North Carolina, Tennessee, Delaware, Florida, Arkansas, Alaska, and Arizona) indicate that growth models may produce little or no difference in AYP ratings from the status models.

There are at least three reasons why growth models implemented so far by the initial group of states in the NCLB pilot program tend not to demonstrate significantly different results than status models. First, on average most students demonstrate about a year's achievement growth every school year (however defined by each state), which means that students who are behind tend not to close the gap very much, if at all. The second reason growth models tend not to differ substantially from status model results under NCLB are the significant restrictions placed on growth models (Goldschmidt & Choi, 2007). The NCLB requirement all schools get all students to proficiency by 2013-14 is a difficult target for a school to meet under either model. The approved state pilot models had to define student growth projections to proficient within three years (a short period of time for many students), and states are not allowed to use confidence intervals for determining growth estimates for a school or student group. The third reason growth models are not having much impact is some states apply growth models only after applying the standard NCLB status and safe harbor methods. Although growth models as they are currently implemented under NCLB may not provide much benefit to states, it should not be inferred that growth models generally cannot provide additional useful information (beyond status models), nor should it be inferred that students in these states are not producing academic growth.

A separate, and quite different, purpose for measuring and reporting growth at the student level is to make the growth estimate an essential component of *program evaluation*. Just as a growth model can provide evidence to monitor student improvement, it can also provide evidence to enhance

the evaluation of programs and make program modifications. Programs (instructional methods, materials and designs) can be judged based on producing growth instead of, or in addition to, producing a given level of achievement at the end of the program. While the same questions about how much growth is valuable, the stakes are lower than when the results are used for accountability. Therefore measuring growth for program evaluation may be more accepted by teachers and administrators as a useful tool in supporting classroom decision-making.

Another alternative for growth is a formative use – for instructional purposes. However, if based on solely on annual statewide assessments, growth will lose much of its potential power as a tool of formative assessment. Such large-scale assessments tend to be insensitive to small differences in instruction because the testing is infrequent and teachers seldom receive results in time to adjust instruction for the students who were tested.

5. Choosing the Right Growth Method (for your situation)

Although there are a large number of possible ways to measure growth and design accountability systems, there are a limited number of methods that underlie those possibilities. The next section describes six methods and eight characteristics that differentiate the models. Table 1 is provided to show how the characteristics vary across the growth models.

5.1 Method Descriptions

5.1.1 Improvement

As we stated above, improvement should usually not be considered a method of measuring growth. It is included here to demonstrate the differences between this method and measure of growth based on tracking the achievement of individual students over time. The change between different groups of students is measured from one year to the next. For example, the percent of fourth graders meeting standard in 2005 may be compared to the percent of fourth graders meeting standard in 2006. The current NCLB “safe harbor” provision is an example of Improvement.

5.1.2 Difference Gain Scores

This is a straightforward method of calculating growth. A student's score at a starting point is subtracted from the same student's score at an ending point. The difference or gain is the measure of an individual's growth. The difference scores can be aggregated to the school or district level to obtain a group growth measure. Growth relative to performance standards can be measured by determining the difference between a student's current score and the score that would meet standard in a set number of years (usually one to three). Dividing the difference by the number of years gives the annual gain needed. A student's actual gain can be compared to the target growth to see if the student is on track to meet standard.

5.1.3. Residual Gain Scores

In this method, students' current scores are adjusted by their prior scores using simple linear regression. Each student has a predicted score based on his or her prior score(s). The difference between predicted and actual scores is the residual gain score and it is an indication of the student's growth compared with others in the group. Residual gains near zero indicate average growth, positive scores indicate greater than average growth and negative sores indicate less than average growth. Residual gain scores can be averaged to obtain a group growth measure. Residual gain scores can be more reliable than difference gain scores, but they are not as easily integrated with performance standards in accountability systems such as NCLB because they focus on relative gain.

There are often two concerns related to using gain scores: the negative relationship between gains and the pre-test and reliability. Calculating gains based on observed scores results in a spurious negative correlation between the pre-test (first test occasion) and gain; however this is not necessarily the true underlying relationship and can be ameliorated by estimating true gain. We discuss reliability and gain scores in the technical notes in Section 10.

5.1.4. Linear Equating

As discussed briefly in Section 3, equating methods set the first two or four moments of the distributions of consecutive years equal. A student’s growth is defined as the student’s score in Year 2 minus the student’s predicted score for Year 2. A student’s predicted score for Year 2 is the score in the distribution at Year 2 that corresponds to the student’s Year 1 score. The linear equating method results in a function that can be applied year to year. If the student’s score is above the expected score, the student is considered to have grown. If the student’s score is below the expected (predicted) score, the student is considered to have regressed. Expected growth is defined as maintaining location in the distribution year to year. One disadvantage is that this method is heavily sample-dependent and results may vary from year to year due solely to changes in the sample of students in the cohorts.

5.1.5. Transition Matrix

This method tracks students’ growth at the performance standard level. A transition matrix is set up with the performance levels (e.g., Does not meet, Meets, Exceeds) for a given year as rows and the performance levels for a later year as columns. Each cell indicates the number or percent of students that moved from year 1 levels to year 2 levels. The diagonal cells indicate students that stayed at the same level, cells below the diagonal show the students that went down one or more levels and the cells above the diagonal show the students that moved to higher performance levels. Transition matrices can be combined to show the progress of students across all tested grades. Transition matrices are a clear presentation of a school's success (or lack thereof) in getting all students to meet standard.

5.1.6. Multi-level

This method simultaneously estimates student-level and group-level (e.g., school or district) growth. There is evidence that multi-level methods can be more accurate than difference or residual gain score methods. However, even though the statistics have been around for many years, only recently has the

computing power, software and expertise been widely available. Therefore, the results of this method appear to be more complex because the methods are still unfamiliar to many people.

5.2. Characteristics of Growth Methods

5.2.1. Database of matched student records over time (Student ID)

Most methods of measuring growth require analysis of individual student's results from two or more years. This means that student records from two different test administrations have to be combined or matched. Until recently, most statewide accountability systems lacked a student ID system that assigned each student a unique identification number that is recorded with any test that student takes as long as he or she is in the system. Without such an ID number, record matching must be based on some combination of name, birth date or other demographic information. Because of changes in that information over time, combining students' test records is usually time consuming and prone to non-matches and mis-matches.

The preferred solution is to develop a student ID system in which the ID number is part of the students' records statewide. This usually means integrating the ID into each school's student information system and maintaining a central database to assign and report the ID numbers. These changes require a significant investment or resources to develop and implement the new procedures. However, in the end there should be a reduction in the work needed to match student records and an improvement in the quality of the information available.

5.2.2. Requires common scale

Some growth methods require student scores to be reported on a common scale in which differences in scores across grades are consistent and meaningful. Ideally, this would mean that all the tests were written with measuring growth in mind and based on content standards that are aligned across grades. However it is possible to create a common scale for existing tests that were designed separately across grades. There are technical issues and controversies about how to do this equating. Psychometric advice from experts should be sought before determining that a set of tests can be combined for measuring growth.

5.2.3. Confidence Interval

A confidence interval (CI) is used to take into account the uncertainty in measurement and measuring growth certainly includes some uncertainty. Sources for uncertainty include the normal measurement error of the test and sampling error. There are well-established statistical techniques for estimating uncertainty, and growth methods use different techniques due to the differences in the way growth is calculated.

Implementing a confidence interval is not simply a matter of applying a statistical technique. A decision must be made about the width of the confidence interval. A typical narrow CI is 68% (or 1 standard error) while a wider CI would be 95% or 99%. If the confidence interval is implemented

around the target for growth, choosing a wider instead of a narrow CI will decrease the chances of incorrectly identifying a student or school as failing to meet the growth target. However, choosing a wide CI also increases the chances of incorrectly stating that adequate growth has been made when in fact it hasn't. Choosing the width of the CI always involves a compromise between those two types of errors. The policy-maker must weigh the consequences of each type of error and choose a CI that best serves the intended purpose of implementing a growth model. Currently CIs are not approved under the ED Growth Model Pilot.

5.2.4. Includes Students with Missing Scores

Student mobility is a potential problem in any method of growth that measures student achievement over time. If large numbers of students (i.e., more than 15%) do not stay in the same school long enough to take the test each time it is administered, then the sample of students whose scores are included in the model may not represent the whole school's enrollment. A problem would arise if the students with missing scores showed significantly higher or lower performance on the test. If student data are missing randomly, then results based on a 25% sample of total will provide very precise and reliable estimates of the whole school (Goldschmidt, Choi, & Martinez, 2003).

In the improvement method, all students' scores are included. However since individual students are not tracked over time, it is possible that the differences in performance of students who are moving in and out of the school contribute to the observed improvement. This could lead to over- or under-estimation of the school's effectiveness. One of the benefits of multi-level models is that they include scores from all students to estimate growth for both individuals and groups even if the student has incomplete data. Models using only two time points require complete data.

A secondary problem with missing scores occurs when some groups have more missing scores than other groups. In that case the lack of data may mean that growth estimates for those groups are less reliable and may have to be excluded from reports. For all methods, the effects of missing scores on growth estimates can be determined and should be examined.

5.2.5. Includes Results from Alternate Tests

Since some methods require measurements on a common scale, if alternative tests (e.g., for students with disabilities, English language learners, or high school end-of-course tests) do not produce scores on that scale, it may not be possible to include those students in the growth calculations. The Transition Matrix method is based on student progress as indicated by changes in the performance levels attained by students. If common performance levels have been set across different tests, the results can be combined. However, meaningful results depend on the assumption that the performance standards were set such that the performance levels on all combined tests indicate that students have the same knowledge and skills.

5.2.6. Growth Question Answered

Growth methods may be distinguished by the questions they answer. Determining the question you want to answer by using a growth method will make it easier to choose a growth method and to interpret the results of that method.

5.2.7. Student Performance Standards Explicitly Included in Definition of Growth

For two growth methods (Linear Equating and Transition Matrix), the performance standard is built into the method. Therefore there is no need to go through a separate process to set standards for adequate growth after the estimates of student growth are obtained. For the other methods, users often conduct a standard setting process similar to the ones used to determine the individual performance standards for students at each grade level.

5.2.8. Handles Non-linear Growth

Some growth methods assume that each student's growth in achievement follows a straight line. This is generally a reasonable assumption. However, there is evidence that growth over many years is curved with elementary grade achievement growing at a greater rate than high school achievement. If growth is measured more frequently than once a year, there may differences in the rate of growth at different times. If you believe that students' growth is nonlinear, it may be necessary to choose a growth method that can statistically model that type of growth.

Table 1: Growth Method Characteristics						
	Improvement	Difference Gain Scores	Residual Gain Scores	Linear Equating	Transition Matrix	Multi-level
Data Requirements						
Database of matched student records over time (Student ID)	N	Y	Y	Y	Y	Y
Requires common scale	N	Y	N	N	N	Y
Psychometric Issues						
Confidence Interval	Independent Groups	t-Test	Model Error Variance	Model Error Variance	NA	Model Error Variance
Includes students with missing scores	Y	N	N	N	N	Y
Includes Results From Alternate Tests (Different scales)	N	N	N	N	Y	N
Growth Question Answered	Did this year's students do better than last year's students?	Is the gain for a group higher or lower than average?	How much growth was produced by a group?	Did students stay at the same percentile ?	Are students in a group making adequate progress across performance levels?	How much of a group's growth is the result of group-level effects?
Student Performance Standards Explicitly Included in Definition of Growth	Y	N	N	N	Y	N
Handles non-linear growth	N	N	Y	N	Y	Y

6. Meeting NCLB Requirements

Although NCLB often refers to student growth, accountability systems that heavily weight measures of growth would probably not have been approved under the rules originally set by the U.S. Department of Education. Now that ED is encouraging options for including growth (see http://www.ed.gov/news/pressreleases/2007/12/12072007.html for the policy as of early 2008), states may want to design their growth models to calculate AYP. The first design decision a state must make is whether to incorporate a growth component into its school accountability system. The second design decision is whether to use a growth model that meets the ED Growth Model specifications. If the state decides it would like to be approved by the ED, then the state must make design decisions about several other aspects, including the nine listed below.

6.1. Number of Years to Reach the Target Proficiency

The state must choose a time frame for measuring growth. Common variations for the Growth Model Pilot include a set number of years (e.g., 3 or 4); a paired grade approach (e.g., by Grade 7 for students whose Start Point was in Grade 3; by Grade 8 for students who Start Point was in Grade 4; by Grade 11 for students Start Points were after Grade 4); or a school-building configuration approach (e.g., by the last grade in the school building, whether the building is K-4, K-5, K-6, 3-5, 4-6, 6-8, etc.).

6.2. Spacing of Intermediate Growth Targets

The state must decide on a method for determining the spacing of growth targets for students each year. Common variations for the Growth Model Pilot include a linear approach, a normed approach, which may or not be linear (the z-score, multilevel modeling, and vertically articulated achievement level examples are all normed or policy-based and not necessarily linear), or a policy value-based approach (Delaware's proposal incorporating Value Tables exemplifies this explicit policy-based approach).

6.3. Inclusion of and Expectations for Students At or Above Proficient

The state must decide how to deal with growth of students at or above proficient who have met the performance standard as measured by a Status approach. Variations include whether to calculate "on track" only for students below proficient, or for all students including those who are currently proficient or above; if calculating growth targets for students who are proficient or above, determine whether an appropriate growth target should be based on their individual growth history, a subgroup average, a state average, or a more complex estimate; and whether to include currently proficient students in the accountability decision based on growth. Note that USDE has not approved systems that give additional credit for moving students beyond proficient.

6.4. Protecting Against Misclassification Due to Measurement Error

The state must decide whether/how to deal with measurement error in the observed score at the Start Point (e.g., by using multiple data for any student estimate) and at any observed score compared to an

Intermediate Growth Target. Variations include using a confidence interval or providing some correction for regression to the mean and other statistical artifacts.

6.5. Protecting Against Misclassification Due to Sampling Error

The state must decide whether/how to deal with sampling error when generalizing from the group of students tested each year to the theoretical population of the school. Variations include using a confidence interval and/or a minimum-n. (See the Technical Note on Confidence Intervals for more information.)

6.6. Dealing with Accountability When Students Change Schools

The state must decide what to do about assigning accountability when a student moves from one school building to another, particularly if the student is performing below a growth target. Variations include making adjustments in the calculation of the growth target, in adjusting the years-to-growth to vary with school configuration, or adjusting the growth target only when a student moves across district boundaries (and not school buildings).

6.7 Dealing with Incomplete Data

Growth models always tend to exclude more students than status models because calculating growth requires at least two years' of data. The state must ensure student inclusion in the growth model, for example through careful student tracking. States may also use the statistical technique of imputation of missing data (e.g., replacing the missing score with a status score or the statewide average). It is recommended that states also have specific plans for monitoring whether the missing data are biased or otherwise impacting the validity of the accountability decisions.

6.8. Reporting

The state must decide at what levels to report results of the growth accountability calculations. Variations include student/subject-area, subgroup [including currently proficient vs. not-yet-proficient], and school. Some states decided only to report the growth accountability results at a school and NCLB subgroup levels, and not to report either assessment results or accountability growth results at the student level.

6.9 Use in Accountability Decisions

The state must decide how to calculate growth. Variations include determining whether each student has met Status-or-Growth or to calculate Status and Growth for each subgroup or school rather than aggregating accountability decisions for individual students. The state must also decide how to incorporate school performance based on growth into the overall school accountability decision. Variations include using the growth determination as a replacement for Safe Harbor, as an addition to Safe Harbor, as a replacement for Status, and as a factor in conjunction with Status/Safe Harbor (e.g., "if Status is at least X and Growth is Y, then the Overall Rating will be Z").

7. Reporting Growth: Guiding Principles

An important part of any coherent assessment and accountability system is the need to effectively communicate the results. Because growth models are new and at times include complex statistical calculations, reports of student growth can be difficult to design. Following are a set of **guiding principles** developed by staff at the Michigan Department of Education (See http://www.ncrel.org/sdrs/areas/issues/methods/assment/as6penc2.htm for the original publication.)

7.1 Accuracy

The quintessential quality required in reporting growth toward attainment of performance standards. Requirements for accuracy apply to the entire spectrum of adopted growth models, from conceptual underpinnings to operational procedures and reporting. Ultimately, fidelity of growth calculations summarized in reports heavily depends on quality checks and attention to detail.

Quality assurance safeguards should be incorporated in all major components of growth model computations, from systematic checking of the student roster file (e.g., file uploads), to third party validation of customized software programs and statistical analyses required to produce growth computations.

7.2 Clarity

Precision without clarity nullifies utility. Reports are certain to become perennial shelf-bound artifacts if readers are not able to quickly comprehend results. If reports are unclear, for reasons related to faulty presentation, esoteric content, or just poor writing, the credibility of the entire growth model effort could be placed in jeopardy. The additional time and effort required in designing visually pleasing and well-written reports can pay dividends. After all, this is what the bulk of the public will typically see.

7.3 Transparency

A close cousin to clarity, this term is used often in growth model circles for good reason. States, school districts and other educational entities can experience a credibility crisis with the public, media, and policy-making bodies, if the model looks and feels like a black box and a credible job is not done to help stakeholders understand the model at some meaningful level.

7.4 Brevity

Being accurate, clear and brief is a great combination that is appreciated by everyone with a busy schedule.

7.5 User-friendliness

An over-used catch phrase in our information age, this is a worthy reminder none-the-less. As mentioned previously, wherever feasible, reports should be designed to be easily grasped, even at-a-glance if possible. The presentation and organization of results should help promote ease of comprehension in spite of the inherent busy-ness of many tabular data. APA style specifications set a respectable standard here. When in doubt, ask your audiences. Presenting prototypes of accountability reports to focus groups can be helpful in soliciting the very kind of feedback you're seeking to ensure user-friendliness. In addition, having graphic artists critique drafts, purchasing resource guides on data analysis presentation or desktop publishing, or even perusing high quality corporate annual reports are additional strategies to consider.

7.6 Comprehensibility

Reading skills at about 8th to 10th, grade level should be about right for most audiences.

7.7 Completeness

It is a balancing act to provide sufficient information and specificity without overwhelming detail. Stopping short of overkill requires knowledge of your audience and the level of comprehension being sought.

7.8 Self-sufficiency

Reports in general, and figures and tables in particular, need to be self-sufficient. Readers, for example, should be able to glean a basic understanding of a chart's content via intelligible titles, variables, and value labels without resorting to reading the longer accompanying text. Explanatory sidebar notes and supporting documentation (e.g., brief glossary of key terms) can make a huge difference in aiding your reader's comprehension without having to seek assistance, which is unlikely to happen in most instances anyway.

8. Growth Models in Action (Examples from States)

A question that quickly arises after deciding to measure growth is, "How much growth is enough?" This is a key question that must be answered before including growth in any accountability system. Most states have had experience with setting standards for individual student proficiency and those experiences can inform the process for setting standards for growth in an accountability system. However there are differences in procedures and in the participants in standards setting. Delaware and Florida provide us with two methods of setting standards for growth in their AYP systems.

Another key question is, "How do we report growth?" There will be more data because there are multiple test scores for each student. New charts and tables will have to be developed and explained to parents, the educational community and policymakers. Hawaii and Michigan describe approaches to analyze, display and report growth being considered in each state.

8.1. Delaware

Growth Targets

To determine how much growth was good enough to make AYP, the NCLB stakeholder group reviewed examples of student performance and the subsequent averages produced from the model. The growth model targets parallel the traditional percent proficient targets. If 100% of the students in a subgroup were scoring at proficient, the growth value for the subgroup would be 300. Therefore, in 2007 the growth target for reading/ELA will be 68% of 300 or 204 and 50% of 300 or 150 for mathematics. The table below shows the targets for both the growth model and the traditional AYP model for reading and mathematics through 2013-2014.

	Growth Model		Traditional Model	
School Year	Reading/ELA	Mathematics	Reading/ELA	Mathematics
2003	na	na	57%	33%
2004	na	na	57%	33%
2005	na	na	62%	41%
2006	186	123	62%	41%
2007	204	150	68%	50%
2008	204	150	68%	50%
2009	219	174	73%	58%
2010	237	201	79%	67%
2011	252	225	84%	75%
2012	267	249	89%	83%
2013	285	276	95%	92%
2014	300	300	100%	100%

The growth model and traditional model calculations will be done by subgroup separately for each content area, reading and math.

Procedures for Calculating Growth

The state has a data system with a unique student identifier that allows for assessment data to be tracked and matched from year to year for each student. The proposed growth model assigns points based on the combination of a student's performance level in two consecutive years.

Grade 2 Level	Grade 3 Level				
	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Below	0	0	0	200	300
Meets	0	0	0	0	300

Year 1 Level	Year 2 Level				
	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

The calculations for the content areas of reading and math are done separately. Points are assigned to the outcomes that are more highly valued by the NCLB stakeholder group. Delaware educators set five levels of performance for reading, writing and math at grades 4, 6, 7, and 9. The grade 2 assessments have fewer items; therefore, three levels of performance were more appropriate than five levels.

Performance below proficiency has been divided into two subcategories to better demonstrate growth below the proficiency level for the growth model. In the “Well Below” category, performance level 1, the performance cut score for the subcategory at each grade level and in each content area was statistically determined to be at the scale score point where the cumulative percentage of students scoring in the well below category was fifty percent (50%). For the “Below the Standard” category, performance level 2, the subcategory was set by dividing the scale score points from the lower bound to the upper bound in half. The levels at or above proficiency, performance levels 3 through 5, are collapsed into one category. The subcategories are only used in the growth model and not used in traditional model including status or safe harbor. Cut scores for reading and math for the growth model are shown in the table below.

Reading Cut Scores for Performance Levels Below Proficiency to Proficiency (PL 3) for Determining Growth					
	PL 1A	PL 1B	PL 2A	PL 2B	PL 3
Grade 2	na	na	na	<337	361
Grade 3	<368	368	387	401	415
Grade 4	<400	400	414	427	440
Grade 5	<413	413	427	440	453
Grade 6	<416	416	435	448	460
Grade 7	<422	422	438	452	465
Grade 8	<448	448	466	481	495
Grade 9	<442	442	468	483	498
Grade 10	<448	448	470	486	501

Mathematics Cut Scores for Performance Levels Below Proficiency to Proficiency (PL 3) for Determining Growth					
	PL 1A	PL 1B	PL 2A	PL 2B	PL 3
Grade 2	na	na	na	<330	351
Grade 3	<363	363	381	394	407
Grade 4	<391	391	408	420	432
Grade 5	<416	416	433	442	451
Grade 6	<434	434	451	459	466
Grade 7	<437	437	459	466	472
Grade 8	<449	449	469	478	487
Grade 9	<467	467	486	500	514
Grade 10	<487	487	506	515	523

Using the value tables, each individual student in the subgroup will earn the corresponding points depending upon the cell in the matrix that equals the growth or non-growth from 2006 performance level to the 2007 performance level. For example, if a student scored in the bottom part of "below the standard," performance level 2a in reading, in 2006 at grade 3 and moved to "meets the standard," performance level 3 in 2007, the subgroup in the school that the student attended in 2007 would receive 300 points. Each student's performance is given a value from the table and the average number of points for the subgroup is calculated. This average growth score is benchmarked against the growth standard set by the NCLB stakeholder group to determine whether the school and district met the growth target. The actual growth is measured against potential growth.

It should be noted that preliminary review of the data show that more than 94% of the students in the state who were enrolled in Delaware public schools in 2005 had a test score on the Delaware Student Testing Program in 2004. The remaining 6% have been included in the traditional model provided they meet the full academic year requirement. Therefore, all students are included in at least the traditional or growth models with 94% included in both models. Further, students who should have been included

but did not participate in the assessment are reflected in the participation rate. Again, the same participation rate is used in both models.

8.2. Florida

Growth Targets

The trajectory benchmarks are built individually for students and separately for reading or mathematics. Therefore, a student will have a trajectory based on their baseline mathematics score and the proficiency cut score for mathematics, which is separate from reading.

Grades and Tests Used for Trajectory Growth and the Percent of Closing Needed Per Year				
Grade Of First Enrollment	Test Used As The Basis For Trajectory	Test Used As Target For Proficiency	Years In Trajectory	Percent Of Difference Closed Per Year
3	3	6	3	33%
4	3	7	3	33%
5	4	8	3	33%
6	5	9	3	33%
7	6	10	3	33%
8	7	10	3	33%
9	8	10	3	33%
10	9	10	2	50%

The following table displays the performance expected of students to be counted as on trajectory for inclusion in the proposed method of comparing school performance to AMO targets.

The Amount of Improvement in Terms of Decrease in the Distance between Baseline Performance and Proficiency Benchmark in the Target Grade	
Year In State-Tested Grade	Decrease From Baseline Assessment In Performance Discrepancy
1	33% of original gap
2	66% of original gap
3	Student must be proficient

Procedures for Calculating Growth

If the total and all subgroups have met the 95% participation target in reading and mathematics, and the total and subgroup have met the other academic indicator (writing and graduation), and the proficiency target has not been met, the process is as follows:

1) Identify if the student has been in membership the full academic year and is tested.
2) Identify the number of years the student has been in the state, using the historic files from the state's accountability system.
3) If the student has been in the state public schools, locate the correct baseline score (using the table above).
4) Based on the student's baseline score and proficiency in the target year, calculate the difference.
5) Compare the decrease in the difference against the Improvement table (above) based on the number of years the student has been in the state.
6) If the student's performance on the current assessment is equal to or better than the minimum from the previous step, include the student in the percent "on track to be proficient" growth calculation to compare against the state's AMO's.

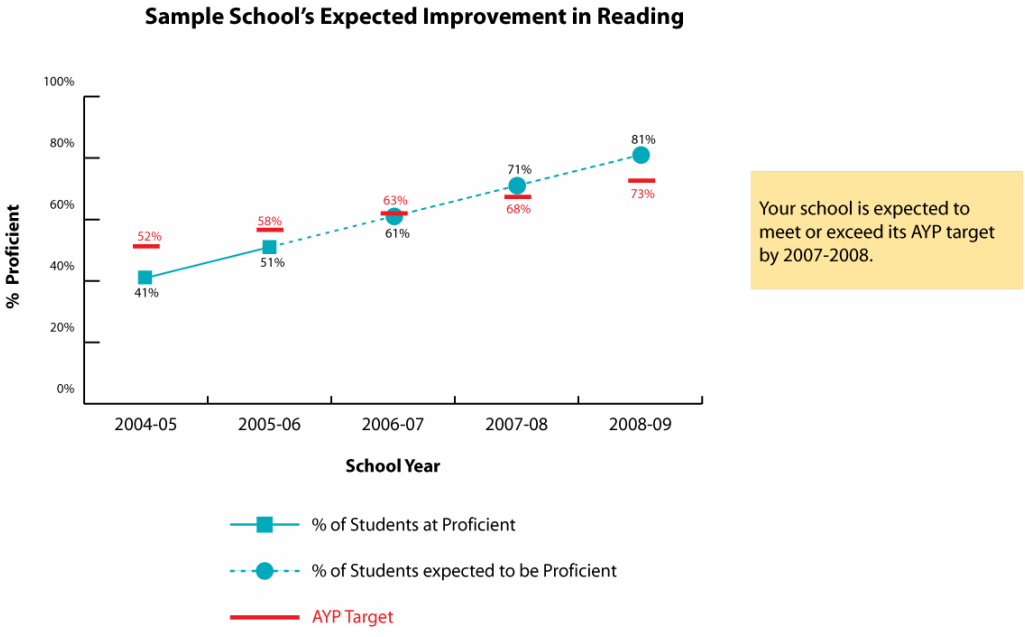
8.3. Hawaii

Background

Hawaii contracted with the American Institutes for Research to help develop a probability-based growth model that computes the likelihood of future proficiency for students and schools, and to develop a report format for displaying such student and school-level results. The group reports use the familiar statistic of percent of students at the proficient level or above. In addition, a projection estimating future performance based on past student growth is plotted. Since targets have been set for future years, the likelihood of a school meeting the expected growth requirements in the future can be estimated.

Reporting Growth at the School and District/State Levels[1]

Reports can visually depict projections of student performance at the school (or similar) aggregate level.



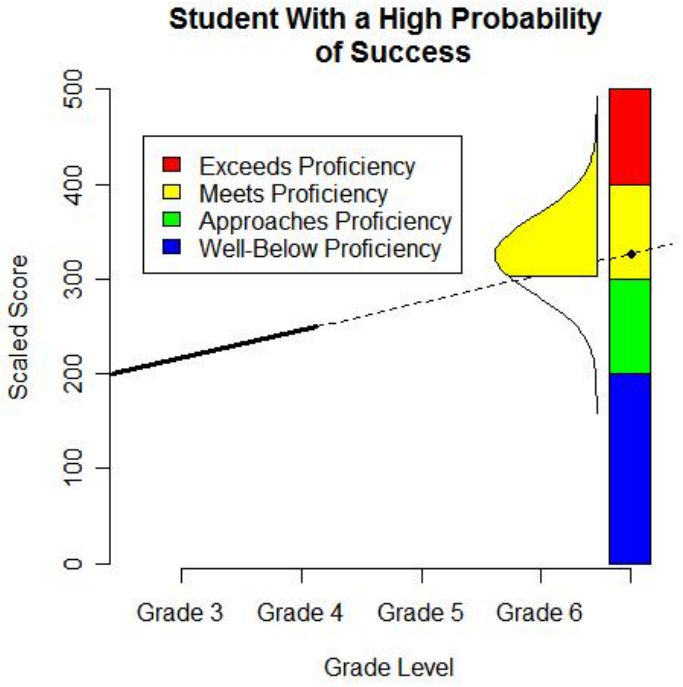
[1] Developed by Harold Doran at the American Institutes for Research

Classroom Level.

Reports could combine brief class listings of individual student growth accountability results with a classroom summary of growth performance depicted graphically over time.

Student Level.

Reports could contain a trend line graphic portraying projected performance in X years to reach Y performance level, based on growth estimates computed for the student. A brief set of explanatory remarks could accompany the graphic and together form the central point of focus for a *Student* or *Family Report*.



Lani had a score of 215 last year and a score of 250 this year. Her score improved by a bit more than the average student's score. About 84 percent of students at this level who are learning at this rate will be proficient by grade 7. Talk to your child's teacher about how you can help maintain or even improve Lani's progress toward the proficient levels. On the following pages, you will find a detailed analysis of Lani's test performance and some specific suggestions that may be helpful.

8.4. Michigan

Background

Michigan developed its first growth reports for the 2006-07 school year. Michigan's assessments are administered in the fall and are reported in the winter so that reports can be used by teachers that still have the students for the rest of the school year. The foundations of Michigan's growth system are:

- The Single Record Student Database keeps Unique Identification Codes that allow matching of student data across assessment cycles;
- Vertically Articulated Performance Standards that enable comparisons of student performance from grade to grade;
- A Value Table approach to growth analysis;
- Reporting on growth to schools, teachers, parents and the public; and
- The use of growth data for school accountability.

Michigan made a key decision to limit its growth analysis to comparisons of student performance at adjacent grades. The rationale behind this decision is the foundation of the system using vertically articulated performance standards. The standard setting process featured concurrent meetings of panelists at adjacent grade levels. Michigan considered that reporting growth across more than adjacent grades was not supported by the scaling, and that domain drift posed problems in comparisons of content and performance across more than one grade level.

Labeling of Performance Change

Michigan also chose to place labels on changes in student performance. The state went through a standard setting procedure, analysis of impact data, revisions to the proposal, analysis policy discussions with stakeholders, discussion with the State Board of Education and a formal public comment period before settling on the following labels for changes in performance:

- Significant Decline
- Decline
- No Change
- Improvement
- Significant Improvement

A value table using these labels is the approved policy instrument. The labels will be used to compare student performance in the fall of 2007 with the same students' performance in the fall of 2006 at the prior grade level. The labels will be used on parent reports, school reports and reports to the public.

The Reporting System

The reporting system gets the data to the point where it can be used. Michigan's growth data are being reported in many ways:

- Parent reports contain the student's performance levels and scale scores for the current year and the prior year. A label describing the change in the student's performance is also provided.

- Teacher reports include class lists containing columns for student's performance levels and scale scores for the current year and the prior year. Labels describing the change in the student's performance are also provided.
- Reports to the public are a "proportions report" showing the percent of students where the change in the student's performance falls into each category.
- The accountability system uses a growth index that summarizes the change in the student performance from the prior year to the current year. The accountability system includes growth data for students only in years can be attributed to that school.

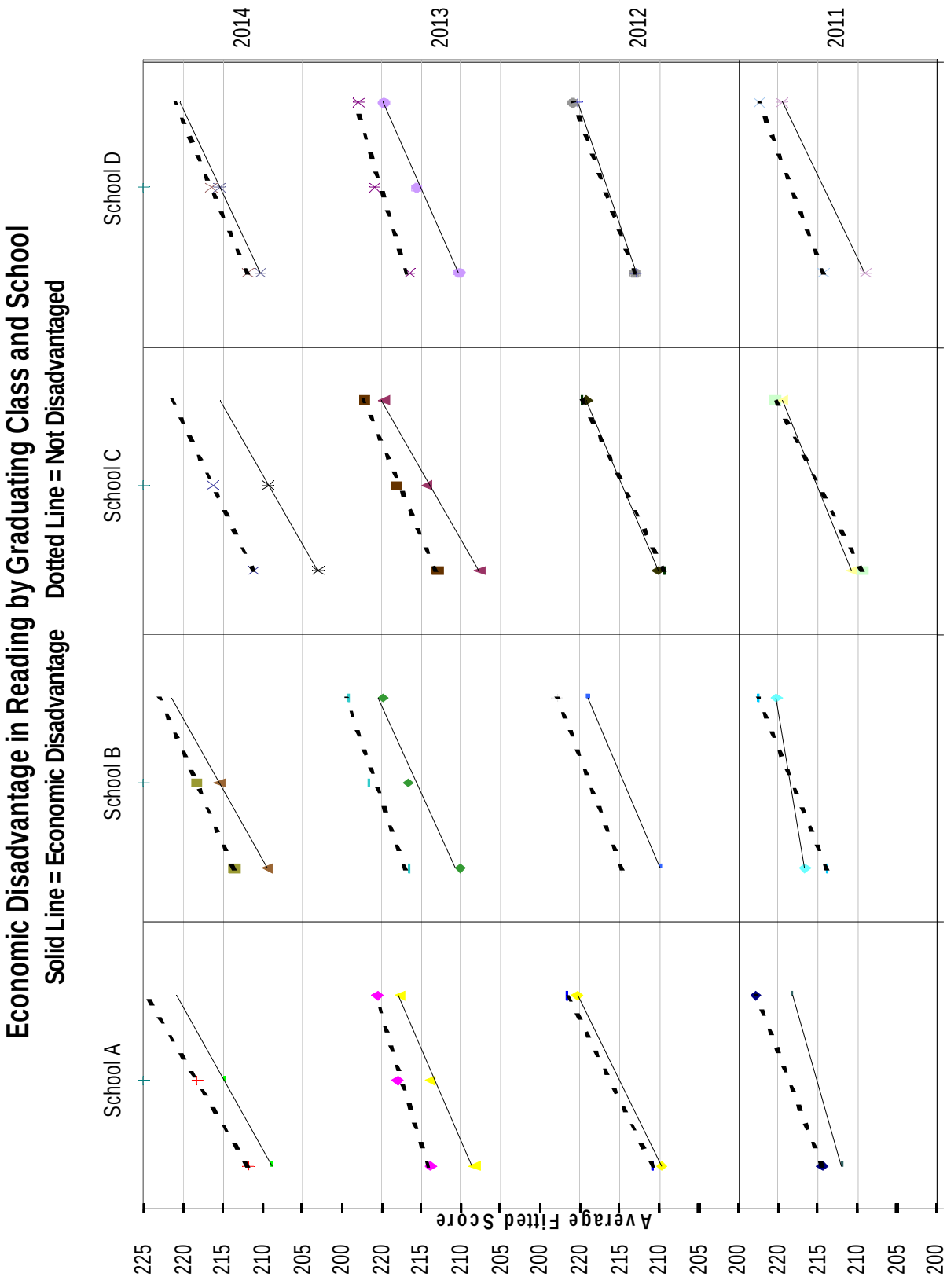
Michigan will implement the growth reporting system in school year 2007-08. It is anticipated that the system will evolve over time, as users ask for data analyses in various formats.

8.5. Oregon's Bridges Project (Reporting on Achievement Gaps)

The Bridges Project is sponsored by the Oregon School Boards Association. The overall goal of the project is to improve student achievement through leadership training for Boards and district administration. One part of the training is designed to improve data based decision-making and growth data has been included. The project uses state achievement data, but it is not a part of the state's reporting system.

One of the uses for growth data that schools involved in the Bridges Project wanted was determining if schools were closing achievement gaps. The lattice chart on the next page shows how the gap in achievement between students with economic disadvantage and students who were not disadvantaged can be displayed. In this case student growth over grades 3, 4 and 5 is plotted by school and graduating class. School staff can compare their results over time and compare their results in any single year with other schools in the district. The results and trend lines were derived from a Hierarchical Linear Model (HLM) and plotted using Microsoft Excel.

This method of plotting gaps in achievement can be used for other groups. However, if there are more than two levels in the group (in this case we had disadvantaged vs. non-disadvantaged) the graph can be too cluttered and therefore difficult to interpret. Similarly, the number of rows and columns should be in the range of two to five to keep the chart readable. Note that straight trend lines are used instead of line segments connecting the actual points. This was done to make it easier for readers to compare charts across schools (horizontally) or across cohorts (vertically). This type of chart can display a lot of information, but users may need some training to use it successfully


Economic Disadvantage in Reading by Graduating Class and School
Solid Line = Economic Disadvantage Dotted Line = Not Disadvantaged
School A
School B
School C
School D
2014
2013
2012
2011
Average Fitted Score
Grade
225
220
215
210
205
200
3
4
5

9. Multiple Indicators of Performance: Incorporating Growth Models

Implementers of growth models must balance policy goals with data availability in order to produce robust results. Robust results depend on resolving both technical issues such as precision, reliability, and stability as well as issues related to the validity of inferences. Growth models can provide substantially more information than status models (Choi, Goldschmidt, & Yamahiro, 2005; Goldschmidt, et al, 2005) but can never the less benefit from considering both multiple analyses of the same data as well as multiple sources of information (Baker, Linn, Herman, Koretz, 2002).

Many state accountability systems, including AYP, present annual, cross-sectional results. For example, a report might include how third grade performance in a school changes from one year to the next. Given that both school improvement and multilevel models represent growth, a natural question might be the extent to which these models' results lead to the same inferences about school performance. This can be addressed using a system that simultaneously models improvement (changes in the subsequent performance of cohorts) and individual student growth.

Table 2, below, summarizes the results of a four level (test occasions, students, cohorts, schools) unconditional growth model. Growth models examining only individual student growth generally find that a substantial majority of the variability in performance lies within schools. In fact based on the data used to generate the results in table one, a model excluding cohort (as a random effect), indicates that about 87% of the variability in student growth is within schools[2]. This implies that only about 13% lies between schools and would be amenable to policies directed at differences between schools. The results of the four level model produce a substantively different picture of school performance than one generated using a status model, a school improvement model, or a traditional growth model alone. The results indicate that the variability in individual growth is evenly split between growth within cohorts and schools and between cohorts within schools. This indicates that much (42%) of the variability between students, within schools, is due to the fact that students are associated with different cohorts. The results also indicate that about half (46%) of the variability in cohort growth (school improvement) is within schools, while the remaining lies between schools. Thus, it is much more likely that explanations of the differences between schools in cohort growth will be accounted for (nearly 55%). Moreover, policies directed at differences between schools likely affect subsequent cohorts, but have much smaller impact on achievement growth of existing students. It is also interesting to note that the variability in initial status is predominantly within schools and cohorts. There is little (7%) variability in status among cohorts within schools. That is schools' student inputs do not change much from year to year.

[2] The results of the three level model without cohort are not presented here, but are available from Pete Goldschmidt (goldschmidt@cse.ucla.edu).

Table 2: Random effects	Variability Breakdown
Between students within cohorts, schools	
Initial Status	84.9%
Individual growth	42.7%
Between cohorts, within schools	
Initial Status	6.7%
Individual growth	42.2%
Cohort growth	45.2%
Between schools	
Initial Status	8.4%
Individual growth	15.1%
Cohort growth	54.8%

Figure 1 presents the relationship among the three indicators of school performance: initial status, cohort growth (school improvement) and panel (individual student) growth. Initial Status is not related to either individual student growth or growth of sequential cohorts. There is a moderate correlation between cohort and panel growth. Plotting pairings of the estimates in a three-panel figure allows stakeholders an opportunity to compare where particular schools rank in terms of both indicators simultaneously.

The first panel of Figure 1 demonstrates how schools compare on initial status (intercept) and individual student growth (grade 2). The "yellow" school, for example, has lower than average initial status, but higher than average individual student growth. The top right panel of figure 1 displays the relationship between school improvement (cyear) and initial status. The yellow school has the same below-average initial status and again has higher than average school improvement. Finally, in the bottom panel of Figure 1, the relationship between school and improvement and individual student growth can be seen. The yellow school appears to be a top performer as it rates highly on both school improvement and individual student growth. In contrast, the green school might be considered a poor performer because it rates below average on both of the growth measures. Of course, under current static NCLB legislation, it is likely that the yellow school would not make AYP, while the green school would make AYP.

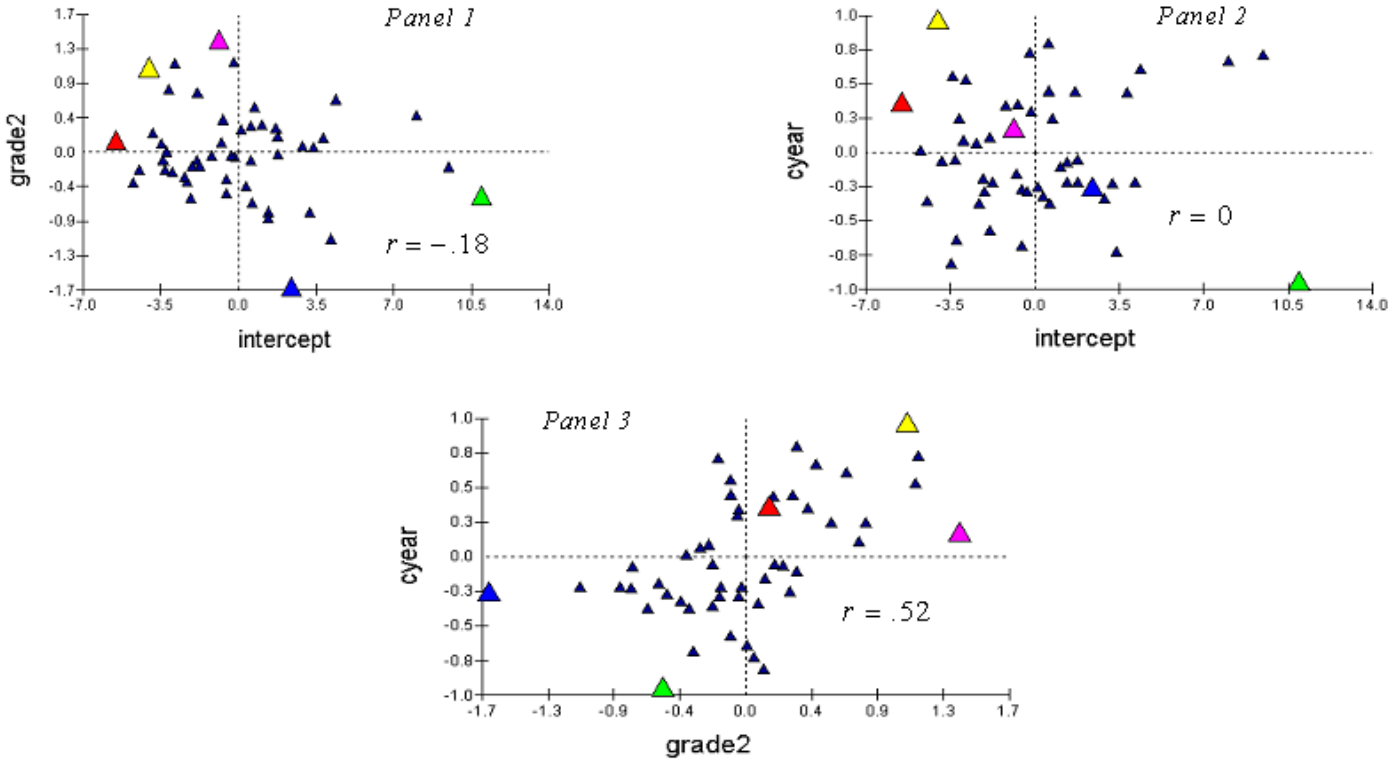


Figure 1: Comparison of school improvement and individual student growth by school.

A growth model that captures both school improvement and individual student growth in this fashion might be the most robust method for monitoring school performance. In addition, beyond matching students from one year to the next (as is required for the more traditional growth models) state data systems can easily provide data to implement this type of model. The disadvantage is that it is technically complex (although reduced two-stage version of this model are possible) both in terms of using the model and in terms of using and explaining results to stakeholders.

10. Technical Notes

We recommend that before delving directly into growth models that states or districts thoroughly understand their assessment and accountability systems. For example, it may be useful to conduct exploratory analyses of data. Plotting average scores by time (by grade) provides an excellent basis understanding what the model is trying to reproduce. This becomes important, for example, when growth is not linear (as discussed below).

Although there is a fortunate correspondence between recent developments in the measurement of growth and the desire for improved methods to hold schools accountable for the real difference they make in student achievement, there are unresolved technical issues related to using growth models. One indication of this problem is that the U.S. Department of Education approved only eight of the desired ten states to use growth in the calculation of AYP after almost two years of reviewing proposals and requiring states to make revisions.

While the issues raised here are related to the technical complexities in measuring growth, they are not without controversy. We decided to include discussion of technical issues on which people may disagree because to exclude them would be to leave out important information for those implementing growth models. Our intention is to inform, not to advocate for a particular course of action. Readers who are uncertain about how to proceed after considering the information provided should seek advice from experts such as members of a state's technical advisory committee (TAC).

As we noted earlier, it is critical to be clear on the **purposes** for using growth. This is good advice for any project, but it is particularly important here because growth must be measured differently for different purposes. For example, if growth is measured using a method that is based on a vertical scale, only those students measured on that scale could be included. This could mean that the achievement growth of students who take an alternative assessment would have to be modeled separately and interpretations of school effectiveness would have to consider this. This is less of a problem for program evaluation, but may rule out that model for use in a high stakes accountability system such as AYP.

10.1. Confidence Intervals

Confidence intervals have been somewhat controversial under NCLB. Some observers believe that defining a school as meeting AYP when it is within a confidence interval (or margin or error) around the standard is actually lowing that standard. The legitimate purpose of a confidence interval is to acknowledge the uncertainty or error in the school's score. There are two relevant sources of error here: measurement and sampling. Measurement error is present in any test and can easily be calculated from the results of a large-scale administration of the test. Sampling error is present when the scores are taken from a subset of the whole population of interest. The sample group may differ from the total population in ways that significantly affect the group score applied to the school. Measurement error and sampling error (when samples are used) should be determined and accounted for in any accountability system.

There are valid professional disagreements about whether and how either type of error should be applied to ratings based on the percent of students meeting a standard (e.g., AYP). The details of that disagreement are beyond the scope of this document. However, it is more important to understand that the issue of using of a confidence interval is only partly technical. In general terms, any decision made when there is any uncertainty can produce false negatives and false positives. Policymakers must also consider the consequence of each type of error. For example, if good things (e.g., additional resources) happen to schools that are identified and those resources are available, then one might want to use a narrow confidence interval or none at all because the consequence of an error of the direction of over-identification is not negative. However, if the consequences of identification are severe and intended for only the extreme cases, then a wider confidence interval may be appropriate.

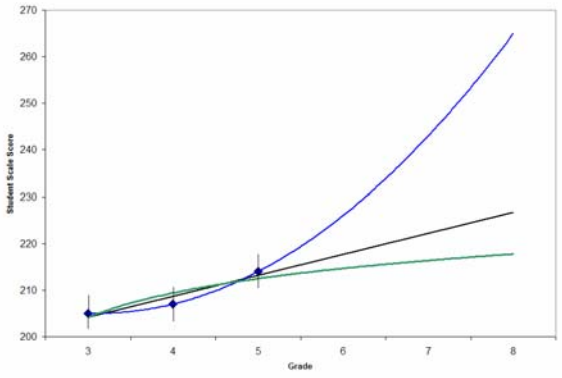
10.2. Expected Growth

In most discussions about growth models, the term of "expected growth" often appears. The reader should be aware of two separate meanings of "expected". Sometimes the word refers to growth that is typical or normal. This can be calculated from the achievement results. The expected growth can be

conditional as in the statement, “The expected growth for ELL students is…” Used in this way, “expected” growth is descriptive of the observed data. The other use of “expected” means “desired” or “required” as in, “Students are expected to...” This is a prescriptive statement.

Since both usages are so common, it is not reasonable to expect that the term be understood as meaning only one thing. Therefore, if the context of the text does not make it clear which meaning is intended, developers of growth models must make sure to avoid confusion.

It is also important to be aware that descriptions of normal growth that extend beyond the period of observed data are based on assumptions that have a significant impact on the rate of growth. An example is charted in this figure (right). A student's three scores in grades three through five are plotted as blue diamonds. Three possible growth projections are plotted. The best fit to those three points is plotted in blue as a quadratic trend line. The linear trend is plotted in brown. A growth projection that takes into account the typical decrease in growth at higher grades is shown in green. It is clear that these three projections vary greatly (over 50 points). Therefore, accountability based on future expectations for student growth (i.e., credit given to “students on track to meet standard”) must determine the validity and reliability of the projections.



When projections are used to draw conclusions about groups of students (e.g., school accountability systems), it is reasonable to assume that the variability will be randomly distributed. In other words, the uncertainty about a student's future score is just as likely to result in an over-estimate as an under-estimate[3]. Therefore, we can be more confident of an accumulated score for a school (e.g., number of students whose expected scores exceed a given target) than we can of an individual's score. In addition, since the amount of uncertainty increases the further into the future growth is projected, one should make projections cautiously. One guideline would be to project future growth by fewer years than the amount of observed data available. In the example provided, there are three years of data (grade 3, 4 and 5) and so projections should be limited to two years or less (i.e., no later than grade 7).

When growth projections are made for individual students, the uncertainty can be calculated and included in a school's accountability rating. Hawaii's report in Section 8.3 provided a good example of how to display the uncertainty of individual students' scores. The student report showed the variability of the expected score using a normal curve and colored bands.

10.3. Reliability and Gain Scores

The advantage of gain scores is that they provide a direct estimate of student growth. Some argue that gain scores are biased and inherently unreliable, but this is not necessarily true (Rogosa and Willet 1983). In fact gain scores can be more reliable than the underlying individual scores. Often the

[3] This assumption will not hold if a high performing group is excluded from the growth calculation, as is the case in some AYP systems, which cannot give credit for student growth above the proficient cutpoint. In these cases, growth will be over-estimated.

unreliability associated with gain scores is due to small sample sizes and lack of variability in gains among students (Raudenbush and Bryk, 2002; Singer and Willett, 2003). It is important to distinguish between true gains and observed gains. Often observed gains are used and these tend to be spuriously (negatively) related to year one scores (Raudenbush and Bryk, 2002). Year to year fluctuations may be too great to provide accurate indicators of school performance (Linn and Haug, 2002).

10.4. Software Packages

When considering the use of growth models, one should be aware that most general-purpose statistical software packages have limitations in the area of modeling growth. Although improvements are always being made, it is advisable to obtain the services of an analyst who has worked with modeling growth for systems that are similar to yours in terms of purpose, size and level of complexity. In addition, states with a technical advisory committee (TAC) should see that the membership includes expertise with recent research and practice in measuring growth.

10.5. Growth Metrics

The issue of the metrics for measuring growth deserves detailed attention. Growth models have potential to more precisely represent the truths we believe live behind the data we use; we expect them to present us with a more convincing, more nuanced story about what works (or not) in schools. However, the interpretation of the results we obtain from growth models can be no stronger than the concepts that the data approximate, nor stronger than the accuracy of the modeling that captures the relationships among the concepts. When about to implement a growth model, it will be wise also to give some thought to what is being measured, how, and why that measurement matters.

We accept some measurement imprecision in testing; we even have name for it, the standard error of measurement. When evaluation of performance was typically relativistic, accomplished by comparing group averages, the imprecision was in fact of limited consequence since error was distributed in a reasonably unbiased fashion.

However, the same cannot be said for statistics that are not parametric. Proficiency indicators as currently constructed are not parametric, nor do they inhabit the kind of conceptual continuum or abstraction that scale scores populate. Rather they are absolute: the student scores above (below) a cutpoint on some distribution of scores; s(h)e is presumed (not) to possess the trait or characteristics that the cutpoint is set to define. The count of students scoring above the cut point (we want to believe) is also the count of students who possess that desired property. However, this assumption should be examined.

First, assume a typical sixth grade reading test with a scale of 100 points, a cutpoint set at 75, and students scoring over much of the scale range. One student scores 75, another 74. The first would be labeled “proficient” while the second is not. If we were to observe these two students in class without seeing their test results, we would probably not be able to determine which was “proficient.”

Second, assume a typical K-8 elementary school. The school’s math proficiency is the sum of proficient students at each grade. Nevertheless, proficiency at grade eight is defined by test items a student at grade three never sees. By definition, proficiency differs by grade.

Third, if the groups are large enough and the comparisons many enough, measures of this form are gross indicators of something. Assume the same test properties and 250 sixth graders. If 40% of the students are found to be proficient, can we be reasonably certain that figure is correct? Probably yes. Assume also there are many more sets of sixth graders. Can we safely say that the group with 40% proficient is performing "better" than a group with 30% proficient? Probably yes. After all, there exist sound and defensible procedures to set cutpoints or performance standards (e.g., Angoff). However, these have very little to say about the substantive meaning of a single student's score (after all, the cutpoint can be attained by correctly answering many different combinations of items and items are unlikely to be fully substitutable). If the error of mis-categorizing a student with respect to the cutpoint is essentially random, the group percentage may still be unbiased and tolerably accurate.

One of the arguments for statistically sophisticated growth models is that they better capture the complexity and reality of life in classrooms. However, if the measures we ask these models to operate with have little discriminatory power because they have been reduced to three or four performance categories, the power of a growth model will have very little to add.

Measurement requires a ruler. A good ruler has at least three important properties. It knows what to count and it has units with which to count. It knows where to start; it has a beginning or origin from which to count. Moreover, it is consistent: it counts the same way every time. A good test is built from carefully selected items that tap some well-defined domain. The number right scores are re-constructed into a scale score that spans the domain, divides that domain into (reasonably) equal intervals, and exhibits consistency from administration to administration. A percent proficient determination, on the other hand, has no scale at the individual level (either the student is proficient or not), no units, and considerable inconsistency. At the group level, there is also no scale: it counts from 0 to 100 consistently enough but that count is of inconsistent individual determinations.

One of the reasons we use a good ruler is that a ruler helps us see when certain thresholds are within reach. That is to say, we expect that behaviors (of student or of teacher) will change when certain events happen, when some mastery or proficiency is reached. Rulers that have only a few values cannot perform this early warning function. Knowing a student is "below basic" is not particularly informative; knowing (s)he is near the top of the category is more useful. However, that requires that the ruler have more values than Yes/No—and that the ruler measure the relevant construct. As Braun and Mislevy (2005) put it, "Effective information about what to do next requires . . . the right items for the right student at the right time. Good information results from good match-ups" (p. 493). Good match-ups can only be determined locally, between teacher and student, between instructional leader and teacher. The best next step is always constrained and supported by current location and present capacity.

That proficiency cutpoints focus school and staff attention on so-called bubble kids, those who score just below a cutpoint, to the exclusion of children already above the threshold or a distance below it, has been documented anecdotally (e.g., Booher-Jennings, 2005). This solution may not be the best next step if we want all students to improve. Neal and Schanzenbach (2007) provide a rigorous demonstration that dichotomous scalings and proficiency percentages (pass rates) have undesirable side-effects on teachers' instructional choices and on students' and teachers' motivation. They conclude that "an accountability system built around threshold scores may not help students who are currently far above or far below these thresholds" (p. 21). For students well below or well above the

thresholds, there is little motivation to put forth additional effort; similarly, the effort cost to teachers to work to increase the performance of these students is too great.

The story that data tell about a school's performance can be very different from the story they tell about a student's performance. Not only the conclusion but also the beginning and the route between the two matter (Seltzer, Choi, & Thum, 2003). All approaches to using data to guide decisions presume a model to take us from the data we have to the decisions we make. Most of today's state and NCLB accountability systems lack explicit models, so the user remains unsure how to connect the dots that lead from data to decision. An accountability system must make it logic model explicit. (This is also referred to as defining a theory of action for accountability.) Then system developers can align purposes, internal statistical processing and data management. Only if the accountability and data logic models are visible, aligned, and understood can schools and teachers determine if these systems and the data flowing within them support or refute their own understandings of how things work and why things change. Understanding the data elements, the scales used, and the metrics that define them is critical.

Growth models that support longitudinal, multivariate, nested analyses of group and individual performance and that attempt to use all available data at once, can more realistically summarize the information available in large collections of imperfect data.[4] Because they identify individual student growth trajectories, these models more adequately control for metric problems and other confounding factors. One advantage is that each student serves as his (or her) own control: history is explicitly accounted for, one student at a time. The overall curvature of a student's growth trajectory becomes a useful index of learning change. Using these methods, out-of-school characteristics of individual students, such as ethnicity or poverty, no longer bias estimates of the gains students make. Rather, the instructional experiences fostered in classrooms and schools become the source of relevant explanatory variables. Given evidence that instructional experiences have differential effects, teachers change their practice.

Growth models make demanding assumptions and enforce strong requirements on both data and users. They attempt refined answers to very specific questions. The functional and logical relationships among data elements tightly constrain logic and inference, method and conclusions. This precision is responsible for their value. The assumptions and requirements form a statistical model. To use these tools, we must accept (or alter) the models and their specifications about how variables are allowed to interact: the statistical models enforce their own rigor about data quality and inference. That rigor must match our mental model of how schooling works. If that match is not present, the calculated results will not be interpretable.

10.6. Units of Time for Growth

A technical issue that seems straight forward, but requires some consideration is time**.** Growth is usually measured as an increase in achievement over some amount of time. Typically, grade level is used as the unit of time. However, it isn't just time that is related to increased achievement. Presumably, the events of teaching and learning are what result in higher scores. At the elementary and

[4] Recent technological advances, Bayesian approaches to iterative algorithms in particular, make it possible to calculate with data arrays that contain a larger variety of disparate measures yet that are tolerant of real-life patterns of missing data (Seltzer, Choi, and Thum, 2003).

middle school levels, an academic year is an adequate indicator of an amount of instruction that is relatively consistent across a school. However, at the high school students may be taking courses that provide more or less instruction related to content on the test. For example, some high schools use a block scheduling to provide longer class periods during the day. Under this system, a student will take fewer classes during a semester and so it is common for a student to take no math during half the school year. A math test administered in May would have a different interpretation for a student who is currently enrolled in a math class as compared with the student who hasn't been in math since January.

Some high schools allow students to take the test when they have completed the classes related to the test content. This may occur over several years. The status measure of percent of student proficient is calculated at a specific time, e.g., end of grade 11. The percent of students who have reached proficiency by that time is the measure of school achievement - even if the student took the test one or two years earlier. While this works as a status measure of individual students and of schools, it may not work as a measure of growth. If the desired measure is the amount of growth produced by the school by grade 11, we are missing the scores of students who took the test earlier. These are the more able students and they would score higher in grade 11 than they did earlier. Therefore, a grade level measure of growth in this high school would underestimate the growth produced by the school. Growth can be measured by grade level for individual students who take the test at different times, but the problem is how to measure the group growth rate across grade levels when the students are not being measured at the same time. This brings up the issue of opportunity to learn (OTL). If students are assessed on content to which they have not been exposed, the assessment results lose validity

One possible solution is to use instructional time as the independent variable for measuring growth. This requires tracking student course taking (transcripts), which are available to high schools, but are not often available at the district or state level. It is also possible to set test administration policy to require all students to be tested at specific times. The problem with this policy is that students for whom the test is no longer relevant because they completed the course work years earlier may not put forth their best effort to show what they know. The issue comes back to the purpose or purposes of the assessment. Measuring growth adds an additional purpose to consider in scheduling large-scale assessments.

11. Suggested Reading

"The Practical Benefits of Growth Models for Accountability and the Limitations under NCLB"
Pete Goldschmidt and Kilchan Choi
CRESST (The National Center for Research on Standards and Student Testing)
Policy Brief # 9 Spring 2007
http://www.cse.ucla.edu/products/policy/cresst_policy9.pdf
This policy brief provides recommendations for going beyond current NCLB definitions of growth models in designing effective accountability systems that include growth along with other factors of school success.

"Growth Models: A guide for informed decision making"
Jim Hull
Center for Public Education
http://www.kintera.org/TR.asp?a=gwKYJjMYIhJ3JuL&s=cnILLKOiEcJMKTMuHoE&m=jvJ0JlMVKeL6F
This guide, put together for the National School Boards Association, summarizes a lot of information on growth modes and provides useful background for district and school decision-makers.

"Putting Education to the Test: A Value-Added Model for California"
Harold C. Doran and Lance T. Izumi, June 2004
Pacific Research Institute
755 Sansome Str. Suite 450
San Francisco, CA 94111
www.pacificresearch.org
The straightforward treatment of growth model and VAM constructs provides a helpful overview of accountability issues and value-added analyses despite the initial intended target audience being limited to California. Importantly, the applied research orientation provides both useful context and innovative ideas to address NCLB reporting requirements that are still relevant today.

12. References

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). Standards for educational accountability systems (CRESST Policy Brief 5). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Booher-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. American Educational Research Journal 42(2, Summer), 231-268.

Braun, H.I. & Mislevy, R. (2005). Intuitive test theory. Phi Delta Kappan 86,7(March), 489-497.

Choi, K., Goldschmidt, P., & Yamashiro, K. (2005). Exploring models of school performance: From theory to practice. In J. L. Herman & E. H. Haertel (Eds.), Uses and misuses of data for educational accountability and improvement (NSSE Yearbook, Vol. 104, Part 2, pp. 119-146). Chicago: National Society for the Study of Education. Distributed by Blackwell Publishing.

Goldschmidt, P. and Choi, K. (2007). The practical benefits of growth models for accountability and the limitations under NCLB (CRESST Policy Brief 9). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). Policymakers Guide to Growth Models for School Accountability: How do Accountability Models Differ? The Council of Chief State School Officers, Washington, DC.

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

Neal, D. & Schanzenbach, D.W. (2007). Left behind by design: Proficiency counts and test-based accountability (NBER working paper 13293). Cambridge, MA: National Bureau of Economic Research.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and Data Analysis

Methods (2nd ed). Newbury Park, CA: Sage Press.

Seltzer, M., Choi, K. & Thum, Y.M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insights into 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.