Using Student Longitudinal Growth Measures for School Accountability Under *No Child Left Behind*: An update to inform design decisions

Brian Gong, Marianne Perie, and Jenn Dunn
Center for Assessment

Executive Summary

Many states are interested in using measures of student longitudinal growth for school accountability, and are considering submitting a proposal for the U.S. Department of Education's (USED) Growth Model Pilot by November 1, 2006. This paper is intended to help states design an NCLB-compliant growth model system. It assumes overall familiarity with NCLB and the states' growth model pilot proposals. The main message of the paper is that there are multiple ways to implement common design decisions for a growth model consistent with the underlying principles of NCLB, and that the states' proposals endorsed by USED illustrate a few ways to implement these design decisions.

The paper provides a summary of design decisions a state should consider in deciding upon a growth model for school accountability. A second section highlights the key design requirement for the USED Growth Model Pilot of "determining enough growth," and analyzes how "enough growth" was handled by the three state proposals—North Carolina, Tennessee, and Arizona—that USED indicated were acceptable in April 2006.

The paper identifies four possible approaches to measuring growth—vertical scales, z-scores, multilevel modeling, and vertically articulated achievement levels. The paper also briefly discusses nine design decisions any growth proposal should address, and which show variation in the approved proposals: number of years to reach the Target Proficiency; spacing of Intermediate Growth Targets; inclusion of and expectations for students at or above Proficient; protecting against misclassification due to measurement error; protecting against misclassification and decision inconsistency due to sampling error; dealing with accountability when students change schools; dealing with incomplete data; reporting; and use of growth decision in overall accountability decision.

Background – *No Child Left Behind* and School Accountability

In the past two decades, elementary and secondary public school reform has been dominated by attention to standards, assessment, and accountability. *Standards* refer to the statements of what students should know and be able to do, and importantly, the policy goal that all students should have access to educational opportunities and instructional supports so they can achieve at least the levels of proficiency established by the state.[1] To help ensure students' attainment of the standards, *assessment* instruments have been developed to measure student performance, and assessment policies have been developed to provide (hopefully) for the valid and reliable measurement of as much of the standards as is practical with a large-scale, publicly funded assessment. Finally, *accountability* policies have been implemented to portray school quality in terms of student performance on assessments and other indicators, and to specify consequences for schools whose students do/do not meet the established performance criteria.

No Child Left Behind (NCLB)—the federal law passed in 2001—gives central place to these elements and overall strategy for school improvement. NCLB specifies that each state must develop content and performance standards in English language arts/reading, mathematics, and eventually science; develop and administer tests aligned to the state's content standards for virtually all students in grades 3-8 and high school; and hold schools accountable for helping their students reach proficiency on the state assessments in English language arts/reading and mathematics. Schools must help increasing proportions of their students score proficient, up to 100% of their students by 2013-14. Between the institution of the NCLB law and 2013-14, schools must meet annual objectives in terms of student performance; these annual objectives apply to all students in the school as well as to racial/ethnic subgroups, economically disadvantaged students, students with disabilities, and English language learners.

Student Longitudinal Growth as a Valid Measure for NCLB School Accountability

Performance of students on standards-aligned state assessments have been used as a primary basis for making school accountability decisions for over a decade, including under NCLB. As described by Dale Carlson and others, school performance might be described in four main ways: a) status, or performance at a point in time without reference to previous performance; b) improvement of successive groups (e.g., grade 3 in 2005 compared to grade 3 in 2004), c) student longitudinal growth (e.g., students' performance in grade 4 in 2005 compared to the performance of the same students in grade 3 in 2004), and d) change in rate of change (either of improvement or growth).

This paper focuses on the notion of measuring the learning growth of students over time, and its value as an indicator of school performance. Validity is—or should be—the heart of all school accountability systems, including NCLB. It is clear from on-going policy debates that many people still wrestle to increase the validity of NCLB, especially in terms of what, who, and how school performance is measured, and the consequences that are specified and implemented.

NCLB dictates evaluating schools on how many students score proficient or above. This is referred to as a *status* measure, because it indicates school performance at a single point in time, namely at the end of testing each year. NCLB also states that a school can

[1] *Standards*, then, include *content standards* (statements of what students should know and be able to do), [individual student] *performance standards* or achievement levels (statements and associated measurement criteria that indicate "how good is 'good enough'"), and *school achievement standards* (statements and criteria of acceptable school performance for accountability purposes).

be considered as "good" if the proportion of students who are proficient increases sufficiently from one year to the next. This is referred to as an *improvement* measure, because it indicates school performance over time of different cohorts of students (e.g., grade 4 in 2005 compared with grade 4 in 2006).

In addition to *status* and *improvement* measures, policy makers, educators, and designers of school accountability systems have discussed *growth* measures as a desirable and valid way to measure school quality. *Growth* measures are based on the learning done by individual students over time, essentially seeking to answer the question, "Did these students increase enough in what they know and can do?" Growth is a more valid measure of schools for many people because it focuses on students' learning more over time, and it is related to schools helping students learn.[2] In contrast, a student can score high on a status measure theoretically without having learned anything in school that year—for example, by coming in at the beginning of the year already proficient. Status measures are typically highly related to the wealth of the students' families/communities. Improvement measures confound differences in scores with differences in cohorts of students from year to year—a "good class/bad class" effect often observed in educational testing. It is conceivable that schools would score differently on these three different measures of performance. For example, a school that scored relatively low on Status might be improving over time (e.g., Grade 4 scored higher in 2006 than Grade 4 did in 2005 or 2004).

Deciding on whether to include Status, Improvement, or Growth in a state's school accountability system depends on what the state values. What does it consider a true indicator of "good" school performance? In addition, states should consider the likely effects of including Growth on the state's "theory of action" that says what it expects and would like to occur as a result of implementing a school accountability system.

While some states pursued incorporating growth measures in school accountability systems prior to 2005, most states did not have the annual testing in adjacent grades or the means to match accurately student test scores to individual students over time, both of which are required for a strong longitudinal student growth system. Prior to NCLB the federal model was to minimize testing, and so most states were testing a sample of students, items, and grades (e.g., once in elementary, once in middle, and once in high school). The NCLB statute specifies that all students must be tested annually in grades 3-8 and at least once in high school. Most states are developing data tracking systems so they have both the required amounts of testing and the usable data to implement growth systems for school accountability. Thus, now is a good time to consider growth models because there is greater conceptual clarity about how they might be incorporated into school accountability systems and it is becoming more practically possible to do so.

[2] For more discussion of Status, Improvement, and Growth school performance measures, see Carlson (2002), available at www.nciea.org under "Publications"; see also Gong (2002), *Designing School Accountability Systems: Towards a framework and process.* Washington, DC: CCSSO, available for download from www.ccsso.org.

Some Key Design Decisions for Incorporating Student Growth into School Accountability

The first and most fundamental accountability design decision a state must make is whether it should include a growth component or not. A state considering student growth should address these key issues listed below in designing their accountability system.

1. What is the purpose for including student growth in a school accountability system?
    - Many educators feel that student growth provides a better dimension for schools to be accountable than status (related to SES) or improvement (subject to "good class/bad class" variation). The idea is that schools should be held accountable for the learning done by the students in the school during a specified time period, e.g., fall to fall.
2. How will student growth be defined, and how will results be reported and used?
    - Student growth is defined as the change in performance (learning) between two (at least) specified points in time. Student growth for the school will be aggregated over all the students for whom the school is accountable for student growth. The aggregation could be a mean (e.g., mean scale score difference) or some weighted average (e.g., as is done with value tables or an index). School performance will be reported in terms of the aggregated score; various disaggregations might also be reported (e.g., by racial/ethnic subgroup, content area, or teacher—although generally these are not recommended for school accountability purposes, although they may be reported for other purposes). The aggregated school growth score will be used in making accountability decisions about the school.
3. How much growth will be "good enough" for school accountability? How will that "good enough" criterion be established?
    - A state must decide whether the growth performance should ensure schools are moving students toward proficiency at an acceptable rate, or whether some other rate and type of growth is "good enough." In general, the current growth rate of student learning in most schools and states would be far below the rate set by NCLB or even the states' own state accountability systems prior to NCLB.
    - A state must decide whether there will be a single "good enough" criterion or multiple criteria for different groups. For example, should students who start lower (e.g., Far Below Proficient) be expected to grow less, more, or the same as students who start higher? Should students who are above Proficient be expected to grow the same amount as other students?
4. How will judgments or ratings about student growth be combined with other judgments (e.g., status, safe harbor) to yield an accountability decision?
    - Will judgment about student growth be compensatory for status and/or improvement/safe harbor, or conjunctive? When would it make sense for it to be mixed, e.g., to make distinctions among levels or consequences, or to be compensatory only under certain conditions?
5. How will the student growth accountability system deal with inclusion issues?

- Measures of student growth require at least scores from two time points on assessments that are comparable. How will the state design its system to ensure maximum appropriate inclusion of students? How will the state deal with students with missing data or who otherwise do not meet the ideal specification (e.g., students retained in grade from the previous year)? How will the state ensure the accuracy and validity of its data used to make judgments about student growth?

6. Are the accountability judgments based on student growth acceptably reliable (i.e., have an acceptable misclassification error rate) and valid?
    - Does information regarding the validity and reliability of the student growth judgments support the intended uses? Was that information obtained in a technically sound way?
7. Does the assessment system support the use of student growth scores in this way?
    - Do the conceptual and operational aspects of the assessment support the measurement, interpretation, and use of student growth scores?
8. How will a student growth system be communicated effectively so the accountability system will have the desired effects?
    - Is there an appropriate balance between sophistication and simplicity? Does the student growth system lend itself to appropriate action?
9. How will student growth be operationalized?
    - Approaches to measure student growth are being implemented by states that use vertical scale scores, vertically moderated achievement levels, and variations of within-grade norming. Statistical treatments range from multi-level, multivariate covariance structures to regression models to weighted counts. The choices about how measurement of student growth is implemented usually reflect decisions about the factors 1-8 above.
10. Is the system sustainable?
    - Are there sufficient resources (time, money, expertise, individual commitment, political will) to make the system successful?

If a state decides that it is interested in including a growth component in its school accountability system, then the state must decide whether it wants its growth model to be USED-approvable and NCLB-compliant.

Growth Models for School Accountability and *NCLB*

There are many reasons to measure "student growth," and many ways to measure growth. This paper considers one specific purpose and one particular set of constraints related to NCLB. For NCLB, the purpose is to provide a measure of schools' progress in helping all students become proficient by 2013-14. Some states are very interested in using growth measures, but do not want to use the same constraints as specified by the USED. Growth should be pursued as a way to increase the validity of the accountability system, not to decrease school identification (unless as a byproduct of more valid accountability) nor to address concerns with other aspects of the NCLB law.[3] States interested in pursuing use of growth models other than those that meet the strict constraints of the USED may try to persuade the USED to change their requirements, or the states may decide to use growth measures not for AYP as a component of a state-only accountability system or only to report results but not for school accountability. Some examples of uses of growth measures that are not acceptable to USED for use with AYP currently: a) determine how much a group of students has grown in relation to the amount of growth achieved by other students, using past performance or student demographic variables as factors; b) determine how much effect a certain program has had on various groups of students (e.g., ELL versus non-ELL); c) determine how much effect a teacher or sequence of teachers has had on a group of students; or d) determine average performances for school, including students both below and above proficient.

The U.S. Department of Education interpreted NCLB as requiring Status and Improvement ("safe harbor") measurements of school performance for accountability, but as not allowing Growth. In November 2005 USED Secretary Margaret Spellings announced a "Growth Model Pilot" program in which up to 10 states would be approved to used student longitudinal data with growth models for school accountability. A primary purpose stated for the pilot was to inform reauthorization of NCLB. The USED solicited proposals from states and established a Peer Review process for reviewing the proposals. Eventually 13 proposals were submitted, 8 of which USED approved for Peer Review. (See Appendix A for a summary of the eight growth model proposals.) USED eventually approved Growth Model proposals from two states: Tennessee and North Carolina. It should be noted that some states were not approved for reasons other than the technical merits of their Growth Model proposals. USED has announced that states may submit growth model proposals in November 2006 (September for the 6 states that were not approved after Peer Review); up to eight additional states may be approved. The USED established several criteria for acceptable growth models, the most important of which that will be discussed in this paper is, "How much growth is enough?"

How Much Growth Is Enough? – NCLB's Policy Goals

A key defining characteristic for any accountability system is determining "how much is enough"? Assessment systems measure performance ("The student scored a 212"), but accountability systems must reflect a judgment about whether the performance is sufficient or insufficient. The USED established clearly that "enough growth" is linked to the same policy goals as the NCLB statute: all students proficient by 2013-14. The ultimate level of performance is proficient on the state assessment; the goal applies to all

[3] Inclusion of growth measures into AYP does not address other NCLB validity issues, including basic assumptions about conjunctive rules; consequences: order, type, scope, quality, effectiveness, barriers, etc.; goals and timeline; invalid assessments; narrowing the curriculum; comparisons across states; other content areas; high school; accountability for excellence, beyond NCLB proficient; special populations: students with disabilities, English language learners, special population schools; influx below tested grades; who's not included: FAY; nor the tension between making valid AND reliable school accountability decisions: Type I and Type II errors.

students; and the time line is by 2013-14. Thus, by design the growth and status measures will converge in 2013-14.

These criteria are somewhat different than have been established when measurement of student growth has been used for purposes of program evaluation and for psychological and academic evaluation. Program evaluation—where the most advanced work has been done with using student growth measures—differs from school accountability. One primary difference is that program evaluation typically considers growth in relation to other groups, not to an absolute standard. Program evaluation also often seeks to control for other variables, and so conditions analyses on background variables such as race/ethnicity, which is usually not acceptable for school accountability.[4]

The USED approved growth model proposals from two states—North Carolina and Tennessee—and indicated it would have approved the proposal from Arizona, pending Arizona's compliance with some issues outside the technical merits of its growth proposal and some particulars in its growth proposal.[5] All three proposals have much in common but also illustrate some important design decisions. It should be emphasized that these three proposals do not illustrate all the technically strong growth models possible, although it is not clear at this point what criteria the USED will apply in reviewing proposals submitted in the fall (Sept./Nov.) 2006.

Basic NCLB Model: Growth to be "On Track" to be Proficient by Target Year

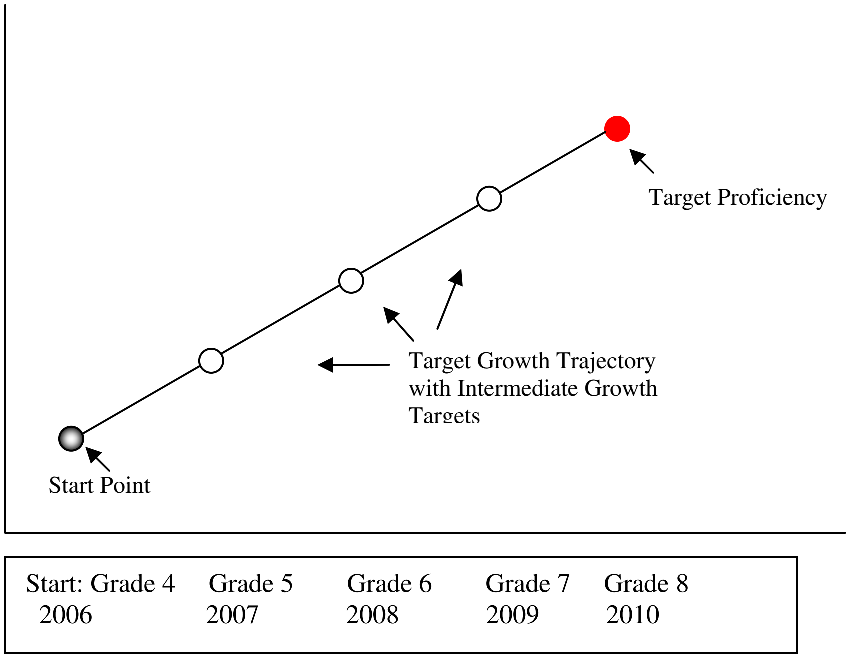
The basic growth model for current NCLB purposes outlines the achievement a student will need to have over a set number of years to move from where s/he is to proficient.

1. Mark where the student is at the Start Point. In our example, the student is somewhat below Proficient in Grade 4 in 2006.
2. Mark in which grade the student is targeted to be Proficient. In our example, the student is to be proficient within four years of the initial baseline year, or by Grade 8 in 2010.
3. Set yearly growth target for the grades/years between the Start Point and Target Proficiency. In our example, circles on the line indicate where the student must score in Grades 5, 6, and 7 to be "on track to be Proficient" by Grade 8 in 2010.

[4] The term "growth models" as used recently in educational measurement circles have referred to statistical and quantitative approaches for measuring same-student performance over time, or for using such data to estimate effects of teachers, schools or other entities. However, "student growth models" also have a long history in developmental psychology where the emphasis has been on characterizing *what* develops and *how* rather than on measuring *how much*. Recent attention to cognitive science and formative assessment has raised the issue of how individual student growth could be assessed more accurately and with greater detail, and that information used to inform instruction as contrasted with holding schools accountable.

[5] While North Carolina and Tennessee were approved, the USED letter to Arizona stated, "As you know, the Department determined that Arizona's growth model proposal seemed poised to meet the seven core principles outlined by Secretary Spellings in her letter on November 21, 2005, and was forwarded to a group of peer reviewers who met on April 17–18, 2006. The peer reviewers indicated that the Arizona model was acceptable provided several changes were made." (Letter from USED Secretary Margaret Spellings to Arizona Superintendent Tom Horne, May 17, 2006. Retrieved from the web on Sept. 13, 2006 at www.ed.gov/admins/lead/account/growthmodel/az/azgmdecltr.doc.)

4. If the student scores at or above the Yearly Growth Target for that grade/year, then s/he is considered to have made enough growth to be “on track to be proficient” in Grade 8 by 2010.
5. The school accountability system credits the school for this student the same as if the student had scored Proficient in that year.



In this basic model, the key is establishing growth targets, and then crediting the school each year that the student’s observed score is equal to or higher than the growth target for that year. The next section describes several ways to establish intermediate growth targets and to link the grades and performances over time.

Variations in Setting Intermediate Growth Targets for the NCLB Growth Model Pilot

For NCLB/USED Growth Model Pilot purposes, the first major design decision involves how the intermediate growth targets are set, which depends on the scale or metric. There are several possible ways to do this, each of which has its own strengths and drawbacks.

Vertical Scale Approach: If all the assessments across the grades share a single common “vertical” or “developmental” scale, then a simple approach is to subtract the scale score of the Start Point (for example, 220) from the scale score corresponding to the Target Proficiency (for example, 280, or 280 minus 220 = 60), and dividing by the number of Intermediate Growth Targets plus one (in our example there will be 3 intermediate growth targets, so 280 minus 220 = 60, divided by 4 = 15 points). The Intermediate Growth Targets in this example would be 235, 250, and 265. Arizona is an example of this common vertical scale approach.

Two strengths to the vertical scale approach are that it can be computationally very simple and the assessment and accountability systems connections can be very straightforward. Potential drawbacks are that a vertical scale is required, which is technically challenging to establish and maintain, and which is controversial in terms of construct validity. In addition, the interval nature of the vertical scale is unlikely to be related in a regular way to the amount of growth needed to achieve standards-based proficiency from grade to grade (e.g., the number of scale score points to grow from Proficient at one grade to Proficient at the next will vary from grade to grade).

Z-Score Scale Approach: If the assessments across grades do not share a common scale, it is possible to generate a common scale, such as a z-scale or a transformation of a z-scale for each grade that allows comparison across grades. Note that the z-score scales must be "frozen" at a reference point in time, against which the growth is then calculated. For example, if a student were at a z-score of 1.2 in Grade 4 in 2005, and Proficient were equivalent to a z-score of 1.4 in Grade 6, then the student would need to grow to 1.3 in Grade 5, and 1.4 in Grade 6. North Carolina is an example of this common z-score scale approach.

Three strengths of the z-score approach are that it can be applied to any set of assessments whether they have a common (vertical) scale or not, z-scores' properties are familiar to many who have worked with NCEs (normal curve equivalents), and z-scores can be transformed into scales that facilitate interpretive focus. Potential drawbacks stem from this application of z-scores being norm-based, including that interpretation of changes over time may be distorted if the current performance distribution is very dissimilar from that of the norming distribution.

Multilevel Modeling Approach: Whether the assessments across grades share a common scale or not, it is possible to use a covariance matrix in a multilevel modeling approach to generate estimates of the student growth trajectories ("slopes") and school effects for each grade/year. By combining the observed scores with the estimated school effects for the (not-yet-observed-for-the-student) intermediate grades/years, one can "project" a student score at the year of the Target Proficiency. By counting the numbers of students projected to be proficient or higher, AYP can be calculated on "projected growth" scores rather than Status. Tennessee is an example of the approach that uses multilevel modeling to produce projected scores.

A strength of the multi-level modeling approach is that sophisticated statistical techniques can maximize the use of available information and minimize some types of error in getting as accurate an indication of student performance as possible. Some drawbacks are that the assumptions underlying the model are usually hidden in a "black box" of complexity few people can understand and thus are difficult to implement, question, or explain. In addition, the particular method of multi-level modeling approved by USED does not draw on the strength of multi-level modeling to separate out effects due to "levels" of performance nested within, or influencing, each other, such as students-within classrooms-within schools-within districts-within state.

Vertically Articulated Achievement Levels Approach: It is possible to use achievement levels rather than a scale. For example, if a student started at Level 1 and Proficiency was Level 3, then the student might have Intermediate Growth Targets of Level 1 in Grade 4 (Start Point), Level 1 Plus in Grade 5, Level 2 in Grade 6, Level 2 Plus, in Grade 7, and Level 3 in Grade 8. Of course, the student is being held to the on-grade-level achievement levels each year, e.g., Achievement Level 2 of Grade 6, Achievement Level 2 Plus of Grade 7. Delaware is an example of a state that proposed using vertically articulated achievement levels.[6]

Two strengths of the vertically articulated achievement levels approach are that it keeps the focus on the state's proficiency standards because the metric is never a scale other than the proficiency levels, and it forces states to deal explicitly with growth between grades, including how much to value growth from all performance levels and how to deal with achievement levels that are not "equal distance" when measured by scale scores (either vertical or z-scores). Potential drawbacks include sensitivity to achievement levels that are not aligned well and less familiarity among the measurement community.

Other Design Decisions

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 USED Growth Model Pilot's specifications. If the state decides it would like to be approved for the USED Growth Model Pilot, then the state must make design decisions about several other aspects, including the nine listed below.

1. Number of years to reach the Target Proficiency – The state must decide how much time it will base its accountability on. 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 7 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.).
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 (the vertical scale example above is linear), 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] In any accountability system, a weight must be assigned to each unit of growth. Most systems use an equal weight for each unit of growth. Delaware combined its vertically moderated achievement levels approach with a method of explicitly assigning weights to each unit of growth (e.g., growth from Level 1 to Level 1 Plus may have gotten more credit than growth from Level 3 to Level 3 Plus). The combination of explicit weights and vertically articulated achievement levels is called a Value Table approach by its developer (Hill et al., 2006. Using value tables to explicitly value student growth. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance*. Maple Grove, MN: JAM Press).

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.
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.
5. Protecting Against Misclassification and Decision Inconsistency 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.[7]
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

[7] The USED Growth Model Pilot Peer Reviewers indicated that they felt "broad confidence intervals" were not technically appropriate for growth systems, essentially since they felt there was not any sampling error. The Peer Reviewers stated, "The justification for employing confidence intervals around the AYP status target is based largely on reducing the impact of score volatility due to changes in the cohorts being assessed from one year to another, and thus reducing the potential for inappropriately concluding that the effectiveness of the school is improving or declining. Under the growth model the issue of successive cohorts is no longer in play since we are measuring the gains over time that are attained by individual students." ("Summary of the Peer Review Team of April 2006," dated May 17, 2006; listed on website as "Cross cutting document." Retrieved from the web on Sept. 13, 2006 at www.ed.gov/admins/lead/account/growthmodel/az/index.html).
This viewpoint that there is no sampling error with longitudinal measurement is incorrect. There is the same sampling error (the "good class, bad class" effect) in trying to generalize from the students who have been tested to all students who will attend the school. The fact that the set of measurements all come from a set (sample) of students, and that every student in the sample is tested does not mean there is no sampling error. This is exactly the same case as testing students and using their scores to make a Status determination. There is sampling error if one wants to generalize from the set of scores obtained to the likely behavior of other students in the school. Every modern school accountability theory-of-action, including NCLB, involves generalizing to future cohorts of students, as is made apparent by examining the prescribed sanctions for schools. The fact that a person measures the same students repeatedly over time and uses the measurements to calculate growth does not eliminate sampling error. For example, suppose we followed a cohort of students who started in grade 3 in 2005, and tested those same students in grade 4 in 2006, in grade 5 2007, and so on. It is clear that it is only one cohort, no matter how many measurements we take. Generalizing to another cohort of students will involve sampling error.

growth target only when a student moves across district boundaries (and not school buildings).

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 decide how to increase student inclusion in the growth model through careful student tracking and through imputation of missing data. Variations for data imputation include replacing the missing score with a status score, the statewide average, or an averaged conditioned score. Some states do not impute missing data but rely on the Status measure for those students; some states also have specific plans for monitoring whether the missing data are biased or otherwise impacting the validity of the accountability decisions.
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 nor accountability growth results at the student level.
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").

Closing Comment

Measuring performance and holding states, districts, and school accountable can help improve student learning. However, we as a nation have put more effort into designing sound ways to measure performance than to improve it. That is true for growth models for school accountability as well. Designing systems to detect growth alone is not enough—educational systems need much more growth than currently is observed, and need desperately to learn how to foster growth. It is not yet clear whether substantial portions of students with disabilities or English language learners can meet high performance standards operationalized in a growth system that requires a student to be proficient within three or four years.

Appendix A: Summary of States' Growth Model Proposals Submitted February 2006

This summary reflects states' final proposals, i.e., those approved by USED for North Carolina and Tennessee.

Tennessee

- School Score = % Proficient or % "On Track to be Proficient"

- "On Track" = proficient in 3 years or less
  - Projected scores
    - Best guess of student's performance three years from now
    - Estimate a trajectory
    - Based on all existing test scores
  - EVASS estimation used for trajectories
    - Assumes Average School Effect

Florida

- School Score = % Proficient + % "On Track"

- "On Track" = proficient in 3 years or less
  - Projected Scores
    - Best guess of student's performance three years from now
    - Trajectory (slope) estimated by subtracting 1st test score from current year test score and dividing by the number of years
    - Uses up to 5 years of test scores
  - Vertical Scale
  - Assumes no School effect

Arizona

- School Score = % Proficient + % "On Track"

- "On Track" = proficient in 3 years or by 8th grade (which ever comes first)
  - Observed Scores
    - Observed performance compared to target
    - Target estimated by calculating the difference between score and proficient score 3 grades from now, divided by the number of years
    - Equal growth intervals
  - Vertical Scale
  - No resetting of target

Arkansas

- School Score = % Proficient + % "On Track"

- “On Track” = proficient in 4 years or less
  - Observed Scores
    - Calculate trajectory required for proficiency (4 years)
    - “On Track” = student makes or exceeds trajectory
    - Largest gains required in 1st year (non equal intervals)
  - Vertical Scale
  - Recalculated every year

Alaska

- School Score = % Proficient + % “On Track”

- “On Track” = proficient in 4 years (for Grade 10, 3 years)
  - Observed Scores
    - Proficient = 300
    - Target = (300 – Year 1 Score) / 4
  - No Vertical Scale
    - All grades have a standard deviation of 75
  - Target reset every year

North Carolina

- School Score = Average Student Growth
  - Student Growth Score = Observed Growth – Expected Growth (Current year test score on z scale)

- Change Score Model (z scale)
  - Proficient Students
    - Expected = Maintaining a positive trajectory
  - Non proficient Students
    - Expected = Proficiency in 4 years
    - Must make up ¼ of the distance each year
  - Expected trajectory does not get reset if student stays in LEA
  - After 4 years, Expected trajectory for non proficient students = trajectory to proficiency in 1 year

- Proposed as alternative decision, but could be used as a growth replacement of AYP

Oregon

- School Score = Average student slope
  - MLM (HLM) used to estimate school slopes
  - Compared to a school growth standard
    - Standard setting procedure
    - Will require an increasing percentage of students to meet growth target each year
    - In 2013-14, all students must meet growth target

- Report % of students that meet growth target
  - Non Proficient Target
    - "On Track" = proficient in 4 years or less
  - Proficient Target
    - "On Track" = Maintain trajectory above proficient
- Vertical Scale
- Growth Score can be used to maintain designation.
- Making adequate growth for two years = making AYP

<u>Delaware</u>

- School Score = Average Value Table Score
- Becoming proficient and maintaining proficient = 300

	Year 2 Level			
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B
Level 1A	25	150	225	250
Level 1B	25	75	175	225
Level 2A	0	25	125	200
Level 2B	0	0	50	125
Level 3	0	0	25	100
Level 4	0	0	0	50
Level 5	0	0	0	0

- Standard = parallel status AMO
- (AMO % of 300)

<u>Summary</u>

- 8 States
  - Add to Proficient Count
    - 2 Change Score (Projected)
      - 1 Vertical Scale
      - 1 Multi-level model
    - 3 Change Score (Observed)
      - 2 Vertical Scale
      - 1 Vertically Moderated (Fixed Standard Deviations)
  - Separate Proficiency Determinations
    - 1 Change Score (Standardized Change Scale)
    - 1 Average Slope (Multi-level Model)
    - 1 Value Table