Exploring Growth Models Under NCLB

Rich Hill, Scott Marion, Center for Assessment
Ellen Forte, edCount, LLC
Bill Erpenbach, WJE Consulting

Goals for this session

- Provide an overview of student growth models for use in state accountability systems
- Discuss the details of the USED requirements for using growth models as part of AYP
- Examine state proposals to gain an understanding of the types of systems likely to be approved for use as part of AYP

Introductions

- Who we are
- Who are you and what do you hope to get out of the session?

- A constraint—We are only going to focus on growth models as they can be used in state accountability systems for AYP.

AYP Reminder

- The following requirements must be incorporated into all state systems used to calculate AYP:
  - All subgroups must be held conjunctively accountable
  - Reading and mathematics must be evaluated separately
  - Students scoring above proficient cannot compensate for students scoring below proficient
  - All schools in the state must be held to the same criteria
  - All students in the state (except AA-AAS) must be held to the same standards and ALL must be included in the system
  - The system must require all students to be proficient or on-track towards proficiency by 2014

Carlson, D. (2002). Focusing State Educational Accountability Systems: Four Methods of Judging School Quality and Progress.

	How good is this school? (Status)	**Is it getting better? (Change)**
Achievement	(A) What is the achievement level of the students in this school?	(B) Is the achievement level of this school improving?
Effectiveness	(C) Is this an effective school? Given the achievement level of students when they enter, how much do they learn or develop while they are in the school?	(D) Is this school becoming *more* effective? How much more, or less, are the students learning this year than they did the year before?

Validity of Accountability Systems

- Many argue, and we agree, that systems where students serve as their own control are more valid—for judging school effectiveness—than status models or cohort improvement systems.

Some Validity Issues with Growth Models

- Metric of Measurement
  - Vertical scale issues
  - Articulated performance standards
- Validity of the gains
  - Compensatory scoring systems
  - Dimensionality concerns
- Validity of the equating
- Normative or standards-based
- Missing data
- The “true” criterion—what are the “right” schools?

The Metric of Measurement

- Almost all measurement models rely on either the use of:
  - vertical score scales (most common)
  - articulated performance/achievement levels across grades (value tables and transition matrices)
- These metrics are crucial to the inferences we make about growth

Vertical Scaling

- Most current growth modeling or value-added models proceed as if there is an equal-interval scale across grades.
- When creating a vertical scale, we need to keep asking if our intended inferences are:
  - Across adjacent grades only
  - Across a significant grade span (3-8)
- The answer will help us focus on the validity of the construct interpretation across the intended span of inference

Validity of gains

- It could also be argued that the vertical equating across grades needs to be validated
  - does a 100 point gain between 4th and 5th grade mean the same thing in each of 2 years?
  - Does a 100 point gain between 3rd and 4th grade mean the same thing as a 100 point gain between 6th and 7th grade?
- How do compensatory scoring approaches affect the validity of our inferences about growth?
- How does the dimensionality of the tests affect inferences about growth?

Validity of Equating

- We and others (e.g., Michaelides & Haertal, 2004; Skorupski, Jodoin, Keller, & Swaminathan, 2003) have become increasingly concerned that many "across year" equating designs are not adequate for capturing change in performance
- For vertical scaling, the validity of the equating must be established both across years and across grades (within years)
- Many growth models based on NRTs have avoided part of this problem because the test remains stable (i.e., equating only once every 7 years or so) and it is measuring—some would argue—a fairly stable trait (e.g., general achievement).

Vertically-articulated standards

- Offers the promise of tying growth to the key tenets of standards-based education
- However, it is dependent on a high quality standard setting process that results in coherent achievement standards
- When the resulting impacts are not identical across grades, it could have a significant effect on the measurement of student growth

The Models

- Many measurement models, ranging from quite simple to very complex, have evolved in an attempt to find the most legitimate and useful ways to capture growth
- Keep in mind, many purposes:
  - Student/teacher feedback
  - Teacher evaluation
  - Program evaluation
  - School accountability for use in AYP—our focus here

A note about data requirements

- At a minimum, the state must have:
  - a method for matching students across grades, usually a unique student identifier
  - a school code and perhaps a teacher or class code
  - demographic information for each student
- The specific file requirements will differ for the particular models
- Missing data/inclusion issue—a growth model will always include fewer students than a status model, but better data systems can minimize this gap.

Types of Models (based on Goldschmidt, et al.)

- Growth Models
  - At the simplest, use the difference in scores between two points in time. Could also use multiple time points. Generally, these are “unconditional” models.
- Value-Added (Residual Growth) Models
  - Rely on sophisticated statistical estimation procedures to explicitly account for accumulation of effects over time as well as student background characteristics. These are referred to as “conditional” models. Many account for (and estimate) teacher effects.
- Hybrid Models
  - Application of one of the models above to include a focus on a clear policy goal (i.e., student proficiency)

Types of Models

- Residual Growth (RG) Models
  - EVAAS
  - HLM
- Growth Models
  - value table
  - transition matrices
  - effect-size
- Hybrid Models
  - REACH model
  - Hybrid Success model

EVAAS (Sanders, et al)

- Complex model using a variation of repeated measures analyses
- Does not assume simple linear growth
- Most common and well-known model
- Adjusts for prior achievement data by including gain scores and includes the effects of multiple prior years of data into current model
- Assumes prior teacher effects remain constant over time
- Requires multiple years of student test scores
- Proprietary estimation procedures requiring intensive computing power
- Can be used to project probability of individual students achieving performance targets

HLM (Raudenbush, Bryk, Choi, Seltzer, Goldschmidt and others)

- A more generalized form of Sanders' model in an HLM framework
- Uses the multiple levels of the analytic framework to estimate the average school growth as well as the distribution of student growth within a school
- Requires at least two scores, but reliability of estimates significantly improve with multiple time points
- Generally, HLM models are conditioned on external variables
- Incorporates measurement error into estimation
- Results are often expressed as residuals and, hence, are normative

Value Tables (Hill, et al.)

- A policy-based model that explicitly values the movement of individual students—aggregated to the school level—across performance levels
- Can be used to establish policy goals
- Requires student-level scores at two points in time
- Not conditioned on anything other than “pretest”
- Assumes/requires that performance standards are coherent across grades

NCLB Value Table--Status

Year 1 Level	Year 2 Level			
	Below Basic	Basic	Proficient	Advanced
Below Basic	0	0	100	100
Basic	0	0	100	100
Proficient	0	0	100	100
Advanced	0	0	100	100

Transition Matrices (Betebenner)

- Based on Markov Chain models used to describe growth as transition probabilities in the language of performance standards
- Can be used to establish policy goals
- Requires at least two time points of data
- No conditioning
- Assumes/requires that performance standards are coherent across grades

Effect Size

- Multiple approaches, but all involve converting difference scores into standard deviation units
  - Used to compare current scores to previous scores (could be baseline scores)
  - Used to set performance goals
- Requires at least two scores
- No conditioning
- Depending on use, tends to report normative results
- Dependent on a somewhat normal distribution of test scores, but not dependent on standard setting or scaling.

Rate of Expected Academic Change (REACH, Doran & Izumi)

- Based on a value added framework
- Uses an individual student's test scores to calculate annual individual improvement targets, based on progress towards proficiency
- Requires two data points, but estimates of progress and goals are improved with multiple years
- Relies on vertical scale, but incorporates scaling error into estimation
- REACH is calculated by dividing the difference between the student’s score and the proficiency cut by the time student is targeted to meet proficiency
- The REACH Ratio is the observed growth divided by the REACH target for the current grade of the student

Hybrid-Success Model (Kingsbury, McCall, Olson, et al.

- Conceptually similar to the REACH model, but does not rely on a VAM framework
- Computes a ratio of actual growth by expected (or targeted) growth
- Based on difference scores using NWEA's RIT scale
- Somewhat normative in that the growth target is dependent on the average growth for students with similar starting positions
- School rating is based on the average individual success index for each student
- Based on pretest-posttest design

Ways that Growth Can Be Incorporated into AYP Plans

- Replace status
  - Compute a growth score for every school and subgroup
  - Establish an AMO
    - 20th percentile?
    - Increasing to what value over time?
  - If school and all subgroups have growth that exceeds the AMO, they make AYP

Ways that Growth Can Be Incorporated into AYP Plans

- Replace Safe Harbor
  - If school or subgroup has status score below AMO, then compute growth score
  - If growth score exceeds some established value, school does not fail AYP because of that subgroup
    - What should the established value be?
    - An increasing value over time?

Ways that Growth Can Be Incorporated into AYP Plans

- Use as a second Safe Harbor statistic
  - If a school or subgroup fails status and status Safe Harbor, then compute growth score
  - If growth score exceeds some established value, school does not fail AYP because of that subgroup
    - What should the established value be?
    - An increasing value over time?

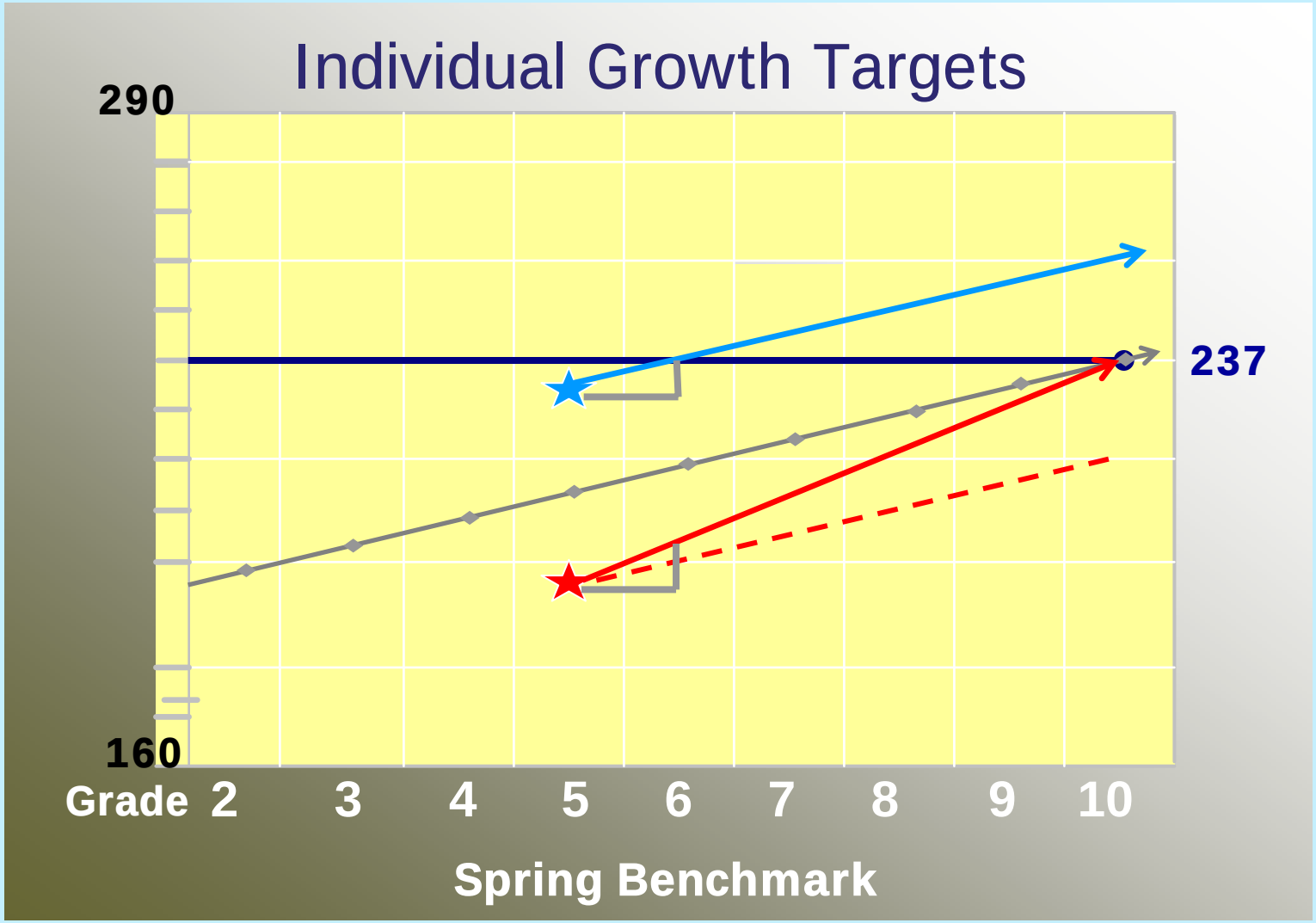
Ways that Growth Can Be Incorporated into AYP Plans

- Determine whether students are “on track” to becoming proficient
- Add these students to the count of those that already are proficient
- Apply usual AYP rules to the new statistic

Determining Whether Students Are "On Track"

- EVAAS, REACH, HGM all establish targets for students
  - Replace percentage of target met with an "all or nothing" statistic
  - EVAAS doesn't use "proficient" as the target, but REACH and HGM do
    - EVAAS projection model calculates the probability for the student to reach proficiency in x years
  - HGM sets targets for students who are above proficient; this would not be necessary

Hybrid Success Model



Vertical Scales Not Necessary

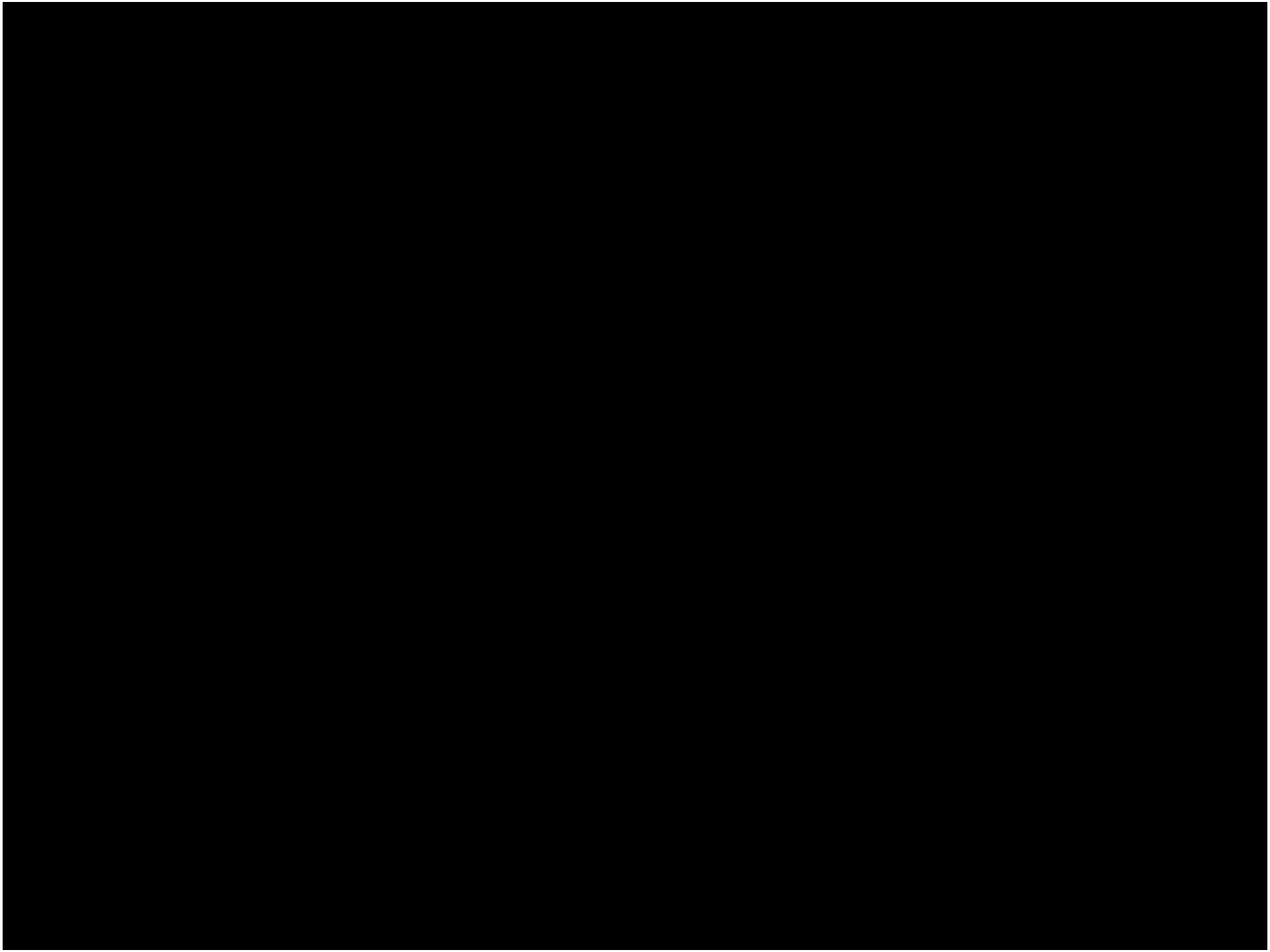
- Articulated standards will do
  - When standards are articulated across grades, can establish equivalent scales across grades by:
    - Fix one point—proficient
    - Make a variance assumption (usually, that variance is equal {or at least, nearly so} across adjacent grades)

Alaska

- At each grade, 300 is the scaled score representing the cut for proficiency, with a standard deviation of 75
- If Year 1 SS is below 300, divide distance between Year 1 and 300 by 4 (or 3, if student was in grade 10)
- If Year 2 SS >= Year 1 SS + result in previous step, student is “on track”

New Hampshire

- Same basic principle, but maximum of 3 years, and students with .5 SD of proficient in one year must be proficient the next in order to be “on track”



Overview

- Pilot program background
- Peer Review Guidance and process
- States' proposals

Background: 2004

- Some states use growth models as part of state accountability systems (i.e., not NCLB AYP)
- Some states have been using growth analyses as part of safe harbor since 2003
- June, 2004: 16 Chief State School Officers wrote to Secretary Paige requesting consideration of growth analyses in lieu of percent proficient in AYP
- Several groups held conferences on the use of growth models for state accountability purposes in the fall of 2004

Background: 2005

- April 7: Secretary Spellings announced a “new approach to implementing flexibility”
  - Consideration of flexibility in other areas if four “bright line” principles are met:
    - Ensuring students are learning
    - Making the school system accountable
    - Ensuring information is accessible and parents have options
    - Improving teacher quality

Background: 2005

- April 7: Secretary Spellings announced a “new approach to implementing flexibility”
  - Consideration of flexibility in other areas if four “bright line” principles are met:
    - Ensuring students are learning
    - Making the school system accountable
    - Ensuring information is accessible and parents have options
    - Improving teacher quality

    (Characterized on May 10 as “annual testing and reporting of student data, plus student achievement and a narrowing of the achievement gap, plus overall sound state education policies, equals a new, common-sense approach to implementation of the law.”

Background: 2005

- April 7: Secretary Spellings announced a “new approach to implementing flexibility”
  - Consideration of flexibility in other areas if four “bright line” principles are met:
    - Ensuring students are learning
    - Making the school system accountable
    - Ensuring information is accessible and parents have options
    - Improving teacher quality
  - “2%” allowance
  - Consideration of growth models
- November 18 and 21: Press Release and Dear Chief letter announcing Growth Model Pilot Program

Background: 2006

- January 25: Peer Review Guidance released
- February 17: States' proposals due
- Up to March 17:
  - negotiations and 19 of 20 proposing states asked to submit additional information
  - seven states chose to defer proposals to 2006-07 school year
- March 31: Eight states selected to move onto peer review phase
- April: Peer review underway
- May: Announcement of states that will be allowed to pilot growth models for 2005-06 AYP calculations

Peer Review: Initial Vetting

Brief proposals first reviewed internally

- State must have:
  - approved or approvable assessment system in grades 3-8 and high school since at least 2004-05 and
  - a data system that tracks student progress
- 19 of 20 states asked to clarify plans or provide additional evidence related to core principles

Proposing States

Peer Review

- Alaska
- Arizona
- Arkansas
- Delaware
- Florida
- North Carolina
- Oregon
- Tennessee

Deferred

- Hawaii
- Maryland
- Nevada
- New Hampshire
- Ohio
- Pennsylvania
- South Dakota

Rejected

- Colorado
- Indiana
- Iowa
- South Carolina
- Utah

Peer Review: Reviewers

- Eric Hanushek, Stanford University
- Chris Schatschneider, Florida State University
- David Francis, University of Houston
- Margaret Goertz, University of Pennsylvania
- Kati Haycock, The Education Trust
- William Taylor, Citizens' Commission on Civil Rights
- Sharon Lewis, Council of the Great City Schools (retired)
- Robert Mendro, Dallas Independent School District
- Jeff Nelhaus, Massachusetts Department of Education
- Mitchell Chester, Ohio Department of Education
- Louis Fabrizio, North Carolina Department of Public Instruction

Peer Review: Core Principles

Core Principles

1. **100% Proficiency by 2014 and Incorporating Decisions about Student Growth into School Accountability**
2. **Establishing Appropriate Growth Targets at the Student Level**
3. **Accountability for Reading/Language Arts and Mathematics Separately**
4. **Inclusion of All Students**
5. **State Assessment System and Methodology**
6. **Tracking Student Progress**
7. **Participation Rates and Additional Academic Indicator**

Peer Review: Core Principles

Core Principles

1. **100% Proficiency by 2014 and Incorporating Decisions about Student Growth into School Accountability**

   - How does the State accountability model hold schools accountable for universal proficiency by 2013-14?
   - Has the State proposed technically and educationally sound criteria for “growth targets” for schools and subgroups?
   - Has the State proposed a technically and educationally sound method of making annual judgments about school performance using growth?
   - Does the State proposed growth model include a relationship between consequences and rate of student growth consistent with Section 1116 of ESEA?

Peer Review: Core Principles

Core Principles

2. **Establishing Appropriate Growth Targets at the Student Level**
   - Has the State proposed a technically and educationally sound method of depicting annual student growth in relation to growth targets?
   - Growth expectations cannot be set or moderated by student demographics or school characteristics.
3. **Accountability for Reading/Language Arts and Mathematics Separately**

Peer Review: Core Principles

Core Principles

4. **Inclusion of All Students**
   - Does the State's growth model proposal address the inclusion of all students, subgroups and schools appropriately?
   - How does the model handle missing data?
   - How are scores from alternate assessments included in the model?
   - How is growth for students in grade 3 (entry to testing) measured?

Peer Review: Core Principles

Core Principles

5. **State Assessment System and Methodology**
   - Has the State designed and implemented a Statewide assessment system that measures all students annually in grades 3-8 and one high school grade in reading/language arts and mathematics in accordance with NCLB requirements for 2005-06, and have the annual assessments been in place since the 2004-05 school year?
   - How will the State report individual student growth to parents?
   - Does the Statewide assessment system produce comparable information on each student as he/she moves from one grade level to the next?
   - Is the Statewide assessment system stable in its design?

Peer Review: Core Principles

Core Principles

6. **Tracking Student Progress**
   - Has the State designed and implemented a technically and educationally sound system for accurately matching student data from one year to the next?
   - Does the State data infrastructure have the capacity to implement the proposed growth model?
7. **Participation Rates and Additional Academic Indicator**
   - Has the State designed and implemented a Statewide accountability system that incorporates the rate of participation as one of the criteria?
   - Does the proposed State growth accountability model incorporate the additional academic indicator?

States' Proposals: Rejections

- Colorado
  - growth targets insufficiently rigorous
  - targets set for cohorts rather than individual students
- Indiana
  - insufficient detail
  - lack of rigor in targets
- Iowa
  - no statewide assessment system
  - bright line principle of teacher quality not met
- South Carolina
  - lack of rigor in targets
  - used growth harbor as a third part of safe harbor
- Utah
  - not reaching 100% in 2013-14
  - combining all non-white subgroups
  - reading and math (and other indicators) combined

States' Proposals: Moved to Peer Review Stage

- Arizona
  - on track to be proficient within three years
  - includes all students in grades 4-8
- Delaware
  - use value tables model after application of classic safe harbor
  - Includes all students grades 3 to 8 and 10
  - proposal appeared to be address the 7 principles most completely
- Florida
  - on track to be proficient within three years
  - includes only students taking the FCAT in grades 4-8 and 11
- Oregon
  - use HLM model for on track to be proficient within four years
  - Includes students in grade 4-8 and 10; not clear if students taking alternate assessments are included
- Tennessee
  - on track to be proficient within three years
  - includes grades 4-8; students taking alternate assessments or with only one year of data are represented in terms of status

For more information

- Scott Marion smarion@nciea.org
- Rich Hill rhill@nciea.org
- Ellen Forte efortefast@hotmail.com
- Bill Erpenbach erpenwj@chorus.net

- www.nciea.org

Table 3
States Identified by ED as not Meeting the Core Principles Required for Growth Models—A Brief Summation of the Identified Limitations

State	Model Type	Principle 1	Principle 2	Other
Colorado	Transitional probabilities—cohorts meet annual proficiency growth targets.	▪Annual growth goals small and could be offset by limits on percentages of students falling back from proficient. ▪Did not reach universal proficiency by 2013-14.	▪Insufficiently rigorous growth targets. ▪Established cohort targets; not individual student targets.	
Indiana	"Longitudinal compensatory" model—students on track to proficiency; used for both status and safe harbor reviews.	▪Detail regarding how schools would be held accountable for growth lacking. ▪Although not cited by ED, model would not reach universal proficiency by 2013-14 as in the case of Colorado. ▪Also not cited by ED, model would serve as a safe harbor review (similar to the case of South Carolina) in addition to a status review.	▪Insufficient detail regarding how growth targets would be established.	▪State's responses to ED's follow up questions on initial submittal provided insufficient detail.
Iowa	To replace status—Students on track to proficiency.			▪Core principles not met. ▪Schools and districts voluntarily adopt State's assessments and not all have participated during past two years; whether all students have participated in the same statewide assessments cannot be established (Principle 5). ▪Issues related to improving teacher quality (a "bright line principle") remain unresolved.

Table 3 (page 2)
States Identified by ED as not Meeting the Core Principles Required for Growth Models—A Brief Summation of the Identified Limitations

State	Model Type	Principle 1	Principle 2	Other
South Carolina	A third screen for safe harbor—Students on track to proficiency.	▪Use of an additional safe harbor calculation does not result in a coherent accountability system. Not clear how growth targets will achieve 100% proficiency by 2013-14 (weighting scheme for proficient and advanced has potential to compensate for lack of growth among lower-performers).		▪Subgroup accountability not maximized because a higher minimum "n" is employed for SWDs and LEP students (Principle 4).
Utah	Value Table	▪Only 75% of students would reach proficiency requirements by 2013-14.		▪Students other than White combined into one large subgroup (other) and no low-income student subgroup for AYP determinations (Principle 4). ▪Reading and mathematics not examined separately; several components combined into one calculation (e.g., attendance, course taking patterns, graduation rates).

Table 2
States Moving to Peer Review—Brief Overview of Selected Components

State	Model	Description	AYP Grades	Students Taking Alternate Assessments	Statistical Tests
Alaska	Count toward AMOs those students proficient and who are on track to be proficient within four years in grades 4-9 and three years at grade 10	A student is on track if he/she is not already proficient and his/her score in the second year is at least as high as the score the previous year plus one-fourth of the gap (1/3 for 10th graders) between the previous year and 300 (proficient).	4 to 8 and 10	Not until that assessment is re-designed.	99% CI
Arizona	Count as proficient those students meeting growth targets toward proficiency within three years or by eighth grade at the latest	Subtract student's current year scale score from the scale score for proficiency three grades higher and divide by the number of remaining grades.	4 to 8	SWDs taking alternate assessment are included; non-proficient students who move up one performance level from one year to the next (there are two levels below proficient) will be considered to have met their growth target (p. 20).	99% CI with subgroups.
Arkansas	Count toward AMOs those students on a pathway to proficiency in four years.	Determine how far a student is from attaining proficiency in literacy and mathematics within a 4-year period and establish a trajectory (pathway) with annual growth requirements. The trajectory will change annually to reflect the student's current scores; potentially re-starting the 4-year pathway. Growth results will also be used to assess which teachers need the most professional development and assistance	4 to 8	SWDs taking alternate assessments and LEP students assessed with a grade-level portfolio are not included.	Uniform averaging and CI; level unknown.

Table 2 **(page 2)**

States Moving to Peer Review—Brief Overview of Selected Components

State	Model	Description	AYP Grades	Students Taking Alternate Assessments	Statistical Tests
Delaware	Values Table Model—Assigns points based on the combination of a student's performance level in two years.	Using a value table, each student in a subgroup earns corresponding points that count toward meeting the growth targets for a given year. Movement from various non-proficient levels to proficient generates increasing values as achievement improves.	3 to 8 and 10	Includes all students.	98% CI
Florida	Students on a growth trajectory to be proficient within the next three years (except for students who will not be in tested grades for three years; they will use the proficiency target for the last tested grade, grade 10).	Growth trajectory slopes are determined by taking the difference between the student's current year FCAT score and the student's first FCAT score and then dividing the difference by the number of years the student has progressed in school since that time (2 to 5 years). Result becomes that student's average annual projected growth rate. Rates are re-calculated annually.	4 to 8 and 11	Not included; only students taking FCAT.	None
North Carolina	Change score scale—used to create growth trajectories for non-proficient students that will bring them to proficient in not more than four years.	Calculate the difference between a student's first test (on the change scale) and the level necessary for proficiency in the grade four years later. The performance target for each year is then based on a 25% decrease in difference per year.	Applies to grades 3-8 but not to high school.	Does not include SWDs assessed against alternate or modified achievement standards.	95% CI

Table 2 **(page 3)**

States Moving to Peer Review—Brief Overview of Selected Components

State	Model	Description	AYP Grades	Students Taking Alternate Assessments	Statistical Tests
Oregon	Multi-level linear model—Tracks growth trajectories for individual students with growth expectations calculated for each student who is already proficient or above as well as for students who are below with the growth expectation set to ensure that the student reaches proficiency within a four-year span.	Standards setting must be completed in order to calculate the performance expectations for individual students and average school slopes are yet to be calculated in order to determine the "kinds of gains that are realistically possible." There will also be a standards setting process to establish annual objectives for school growth.	4 to 8 and 10	Unclear whether SWDs taking alternate assessments are included or whether students taking assessments with accommodations are included.	A statistical confidence interval around the school average growth or slope will be used.
Tennessee	Projection Model—use individual student achievement data to project likelihood of students to attain proficiency on State assessments three years into the future.	Using past score data, apply the State's projection methodology to estimate an individual student's academic achievement level (based on TCAP) at some point in the future (6th grade for a 3rd grader).	4-8; 3rd grade included only in terms of proficient/ non-proficient.	SWDs taking alternate assessments, and any other students without two or more years participation in the TCAP assessments are included in growth calculations only in terms of their current year proficiency scores (as a practical matter, they can not be included because they do not have achievement projections.	95% CI