



## Through-Course Summative Assessments:

### Measurement Challenges and Potential Approaches

March 10, 2011

## RTTT Requirements for Comprehensive Assessment Systems



- Build upon **shared standards** in mathematics and English language arts for college- and career-readiness;
- Measure **individual growth** as well as proficiency;
- Measure the extent to which each student is on track, at each grade level tested, toward **college or career readiness** by the time of high school completion and;
- Provide **useful information** to inform:
  - Teaching, learning, and program improvement;
  - Determinations of school effectiveness;
  - Determinations of principal and teacher effectiveness for use in evaluations and the provision of support to teachers and principals; and
  - Determinations of individual student college and career readiness, such as determinations made for high school exit decisions, college course placement to credit-bearing classes, or college entrance.

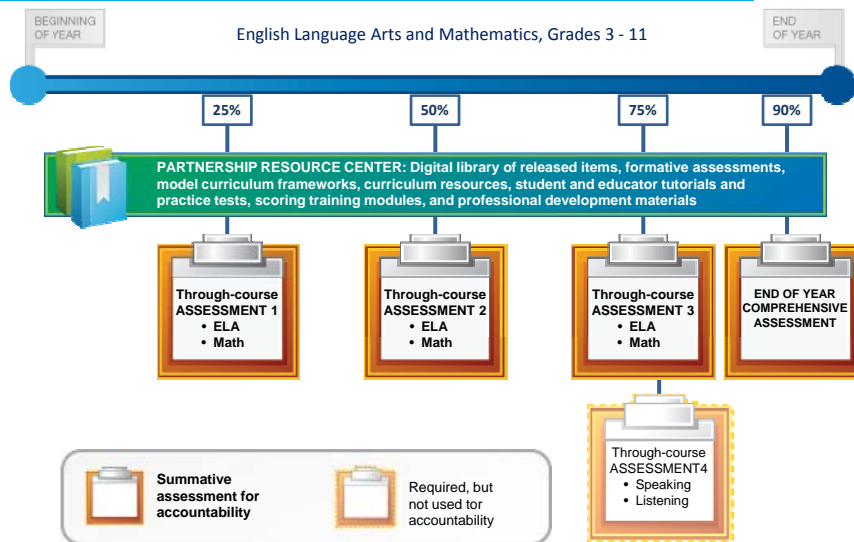
## Innovative Design Feature within RTTT Assessment Program Application



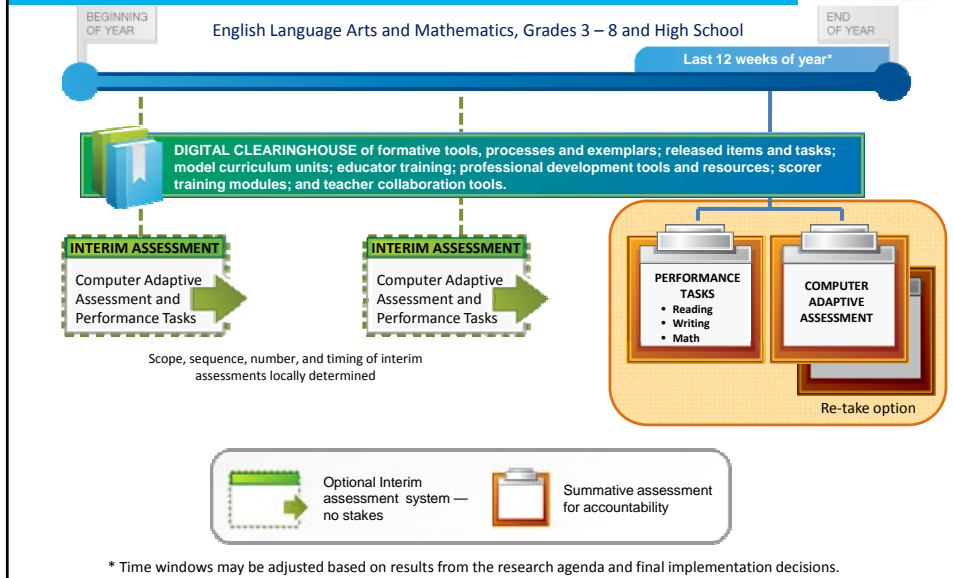
“Through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the academic year. A student’s results from through-course summative assessments must be combined to produce the student’s total summative assessment score for that academic year.”

(US Department of Education, 2010)

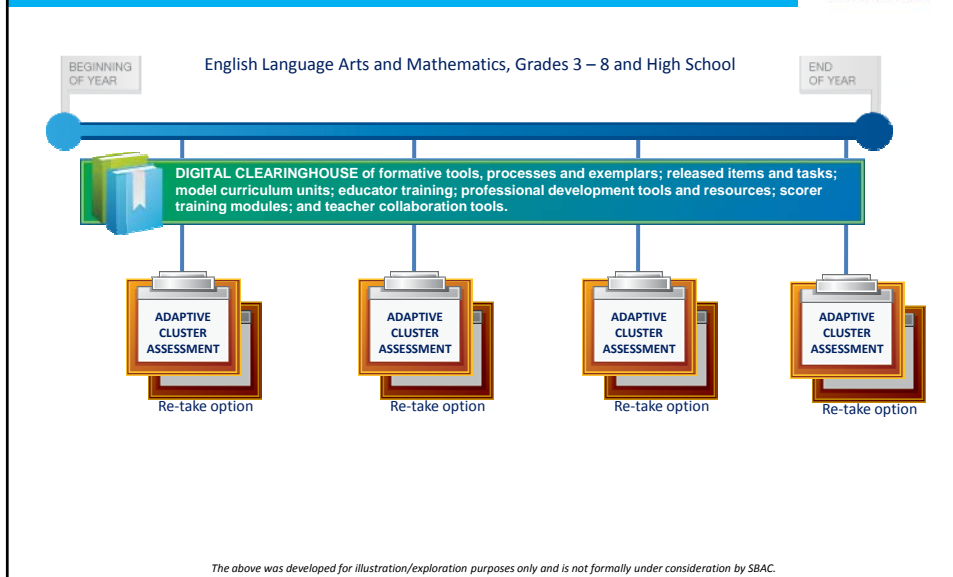
## Partnership for the Assessment of Readiness for College and Careers (PARCC)



# Smarter Balanced Assessment System (SBAC)



# One Possible Model for an SBAC Alternative Summative Assessment



# Measurement Challenges




- A. What **theory of action** underlies through course assessment in the RTTT proposals and how might we evaluate the **validity** of those assessment systems in the context of that theory.
- B. What are the most promising approaches to the **aggregation of results** and the benefits and trade-offs of each?
- C. What are the most promising approaches consortia can use to address the need for **generalizability and reliability** of results from the through-course components, including performance-based tasks?
- D. What are the **essential skills and concepts** that should be emphasized in through-course components, and must they be delivered in a specific sequence?
- E. How can through-course components be placed on a **common scale and linked** from year to year?
- F. How might through-course assessments be used in the **measurement of student growth** and “on track to college/career readiness”? What are the options for incorporating these components into a growth system, and what are the trade-offs?

# Webinar Agenda

Through-Course Summative Assessments (TCSAs)



- |   |                                     |
|---|-------------------------------------|
| 1. <b>Scaling, Linking and Reporting of TCSAs</b>       | <b>Rebecca Zwick, ETS</b>           |
| <i>Questions and Answers</i>                            |                                     |
| 2. <b>Aggregation of Results of TCSAs</b>               | <b>Laurie Wise, HumRRO</b>          |
| <i>Questions and Answers</i>                            |                                     |
| 3. <b>Supporting Growth Interpretations Using TCSAs</b> | <b>Andrew Ho, Harvard GSE</b>       |
| 4. <b>Commentary on Growth</b>                          | <b>Henry Braun, Boston College</b>  |
| <i>Questions and Answers</i>                            |                                     |
| 5. <b>Alligators and Opportunities</b>                  | <b>Bob Rothman and Pat Forgione</b> |
| 6. <i>Questions and Answers</i>                         |                                     |




Listening. Learning. Leading.™

# A Model for Scaling, Linking, and Reporting Through-Course Summative Assessments

Rebecca Zwick  
Robert J. Mislevy  
Educational Testing Service  
March 10, 2011 Webinar

Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved.

3/8/2011



Listening. Learning. Leading.™

## Overview of Presentation

- General model for scaling, linking, reporting
- Example of model application
- Possible simplifications
- Recommendations

2 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved.

3/8/2011

ETS Listening. Learning. Leading.™

## Inferential Goals for TCSAs

- At each occasion, estimate individual student proficiencies and proficiency **distributions** (not just means) for student groups
- Provide end-of year summaries of all results.
- Measure growth across the academic year
- Provide results that are comparable across classrooms, schools, districts, & states


3 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

ETS Listening. Learning. Leading.™

## A complex and flexible model is needed to support these inferences: The model must:

- Provide a structure for linking across occasions and test forms
- Be multidimensional to accommodate multiple subareas (and to satisfy IRT conditional independence – see paper)
- Incorporate items that vary in complexity and instructional sensitivity
- Accommodate varying curricular backgrounds.

4 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011



Listening. Learning. Leading.™


## Proposed scaling, linking, and reporting model

- Builds on large-scale assessment experience (NAEP; PISA, TIMSS)

Unlike NAEP:

- Incorporates market-basket reporting approach
- Yields individual (as well as group) estimates

5 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011




Listening. Learning. Leading.™

## Model includes:

- A multidimensional item response theory (MIRT) model that specifies the dependency of item responses on proficiency.
  - Performance items can be accommodated.
- A population component that models the association between proficiency and background variables (via latent variable regression).

6 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011




Listening. Learning. Leading.™

## Bayesian framework (Mislevy, 1985)

- Item response model & population model are combined to estimate *posterior* proficiency distributions.
- For group estimates, background variables are incorporated in the population model to improve precision and avoid biases.
- Curricular variables are included, allowing the model to reflect students' varying curricular background.

7 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011



Listening. Learning. Leading.™

## Individual proficiency estimation

- To estimate individual proficiency, use mode or mean of individual's posterior distribution.
- For individual proficiency estimates, background variables are **not** used (unless predicting future performance – more later).
- Group characteristics are NOT estimated using aggregates of individual estimates

8 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

ETS Listening. Learning. Leading.™

## Market-basket reporting (Mislevy, 2003)

- Reporting model can be simpler than analysis model.
- Report results in terms of a scale based on a “market-basket” of selected tasks (must be calibrated).
- Using observed item responses, we can impute responses to these MB tasks
- We can weight these responses as we choose.

9 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

ETS Listening. Learning. Leading.™

## Market-basket (MB) reporting (cont.)

- Projecting results onto MB scale is a way to link across multiple forms and over time.
- Multivariate proficiency estimate is mapped onto unidimensional scale.
- “Behind-the-scenes” machinery is complex, but resulting scores “look like” ordinary test scores.
- MB could include administered items, as in forthcoming example:

10 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

ETS Listening. Learning. Leading.™

## Simplified example- Grade 3 Math

- Math is defined as Numerical Operations (20%), Algebra (40%), Geometry (40%)
- Assume each item measures only one area (not a general model requirement)
- 4 TCSAs, each with 30 items (10 per area)
- Difficulty depends on what was just taught


11 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

ETS Listening. Learning. Leading.™

## Example (cont.)

- The  $4 \times 30 = 120$  items are the “market basket”
- For each TCSA, each student has actual item responses for 30 items and imputed responses for 90 items
- Score is always the expectation of importance-weighted sum of responses over all 120 items
- Resulting score has a maximum of 120; provides a metric for linking all 4 TCSAs


12 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

 Listening. Learning. Leading.™

## Market basket scores can serve as the basis for:

- Year-end summary of TCSAs (with desired weights)
- Growth measurement (e.g., last TCSA score minus first TCSA score)
- Projection of end-of-year performance: Given his observed item responses, how would Johnny perform if he had a whole year of instruction (requires curricular variables)?

13 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011

 Listening. Learning. Leading.™


## A possible simpler (and probably cheaper and faster) approach

Use a more traditional assessment (no complex performance tasks) for comparisons across schools, districts and states:

- Administer special booklets (parallel forms with machine-scoreable items) to a random sample.
- If sufficiently reliable, could eliminate the need for separate population model; i.e., individual results could be combined to get group results.

Use less constrained test forms, including complex tasks, in the classroom, to inform instruction.

14 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved. 3/8/2011



Listening. Learning. Leading.™

## Recommendations

1. Use pilot and field test periods to test model and explore simplifications
2. Recognize that a tradeoff exists between inferential demands and procedural simplicity.

- **Reducing demands makes simpler approaches more feasible.**

15 | Confidential and Proprietary. Copyright © 2010 Educational Testing Service. All rights reserved.

3/8/2011

# Aggregating Results from Through-Course Assessments

*Presenter:  
Lauress L. Wise*

*Presented to:  
ETS Webinar on  
Through-Course Summative Assessments  
March 10, 2011*

**HumRRO**  
Human Resources Research Organization

## America's Wake-Up Call

- “Jobless recovery” from current economic crisis highlights the reality of global competition.
  - For jobs as well as goods and services
  - TIMSS and PISA results suggest foreign labor forces are not just cheaper, but **actually better educated!**
- States have adopted a set of common student achievement goals with surprising rapidity.
  - Leading to high standards of readiness for college and careers
- Two consortia are building common assessments of student progress in meeting these standards.
  - Providing us feedback on whether students are receiving the education they need and deserve

**HumRRO**  
Human Resources Research Organization

2

## *Both Consortia Plan Through-Course Assessments*

- The Partnership for Assessing Readiness for College and Career (PARCC) Consortium is considering:
  - 3 quarterly assessments
    - The first two with one or two tasks assessing key knowledge and skills
    - The third would extend over several class periods and cover skills that are hard to measure with short answer questions.
  - A final comprehensive assessment
- The Smarter Balanced Assessment Consortium (SBAC) is considering:
  - 3-4 adaptive tests covering different content
  - Passing each part could substitute for passing the comprehensive assessment.

## *Aggregation of Scores Across Through-Course Assessments*

- The PARCC Consortium intends to combine results from the through-course assessments into summative scores to be used for accountability.
- This presentation explores different methods for aggregating scores and the conditions under which each method provides accurate estimates of student mastery of grade-level goals.

## *Potential Value of Through-Course Assessment*

- **More accurate (reliable) overall results**
  - Through increased testing across multiple sessions
  - Through closer matching of assessment to instruction
- **More timely information**
  - Allowing for mid-year adjustment to instruction
    - Overall and for individual students
  - Providing quicker feedback on effectiveness of curriculum and instruction
- **More diagnostic information**
  - Through deeper measures of subsets of key knowledge and skills

## *Concerns about Through-Course Assessments*

- **Tooooo much testing!**
  - May need some time for instruction
- **Mid-year tests may force a common curriculum!!**
  - Concern that implications for instructional sequences are unproven
  - Many states may be unwilling to sign on to a common curriculum.
- **Mid-year results may underestimate the impact of a full year of instruction.**
  - And they will if aggregation methods are not appropriate to how students learn during the year.

## *Models of How Students Learn*

- Modeling student learning is essential to evaluating different approaches to aggregating test results.
- Results are reported here for 4 general models of how students learn the material being assessed.
  - **One-Time Learning**
    - Students master a topic when it is taught, not before, not after.
  - **One-Time Learning with Forgetting**
    - Students master a topic when taught, but may later lose mastery
  - **One-Time Learning with Reinforcement**
    - Student mastery increases after initial instruction through reinforcement by subsequent instruction
  - **Continuous Learning**
    - Student mastery of a key skill increases somewhat linearly throughout the year.

## *Aggregation Methods Examines*

- Scores from a single end-of-year assessment are compared to the following aggregate scores:
  - **Average Score**
    - Overall results are the average (or sum) of scores from each quarterly assessment (no alignment with instruction assumed).
  - **Maximum Score**
    - The highest score from multiple opportunities to take the same test is used
  - **Matched Score**
    - Quarterly assessment content is matched to the material covered by instruction that quarter; scores are averaged.
  - **Projected Score**
    - Scores from each quarter are adjusted to estimate a full year of growth; a weighted average of these quarterly estimates is used.

## *An Ugly Truth about the Measurement of Growth*

- Individual student test results contain significant measurement error.
  - A 95% reliability means that 5% of the observed score variance is error variance.
  - The standard error of measurement (SEM = square root of error variance) is .22 standard deviations in this case.
- Difference scores used to assess growth contain even more measurement error.
  - Assuming independence, error variances add.
  - .95 reliability on each tests lead to an SEM of .32 s.d. for difference scores (Annual growth may only be .33 s.d.).
- Other (e.g., regression) methods for assessing growth do not greatly improve accuracy.

## *Simulation Study*

- A simulation study was conducted to illustrate the impact of matching different aggregation methods with different models of how students learn.
- “True” growth scores were generated for 400,000 simulated students under each of the four learning models.
  - Average annual growth was set to 1.0 for each model.
  - A standard deviation of .61 was used for illustration.
    - About 5 percent of students would have zero or negative growth.
- Simulated quarterly assessment scores were generated for each of the 4 aggregation methods.
  - Assumed reliability of .90 for quarterly measures
  - Compared to a single EOY test with a reliability of .95

## *Simulated Average Quarterly Growth*

- Average cumulative growth at the end of each quarter under each of the student learning models:
  - **One-Time Learning (1TL) Model**
    - 1.0 in quarter of instruction and each subsequent quarter
    - 0.0 in quarters prior to instruction
  - **One-Time Learning with Forgetting (1TL/F)**
    - 1.15 in quarter of instruction
    - Decreasing .10 in each subsequent quarter
  - **One-Time Learning with Reinforcement (1TL/R)**
    - 0.85 in quarter of instruction
    - Increasing 0.10 in each subsequent quarter
  - **Continuous Learning (CL) Model**
    - 0.25 in the first quarter
    - increasing 0.25 in each subsequent quarter

## *Simulated Growth Averaged Across Quarter Taught*

- Average gains are .25 each quarter for the Continuous Learning (Cont.) and 1TL model
- Early average gains are a little more for the 1TL/F and a little less for the 1TL/R model.

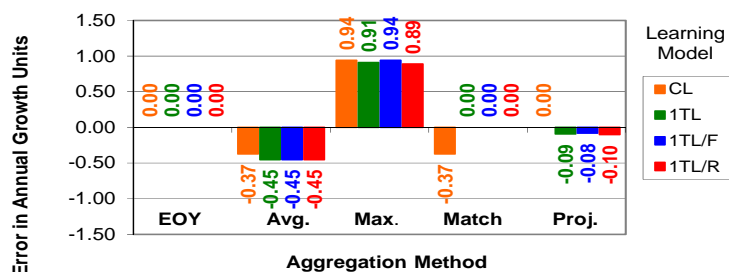
## Simulation Results

- Simulated scores were aggregated using each method and compared to true annual growth.
- Results were also compared to “maximum” annual growth (highest of cumulative quarterly growth values).
- For the “Projection” Method:
  - Quarterly growth scores adjusted to estimate annual growth (x4 for Q1, x2 for Q2, x1.33 for Q3; X1.0 for Q4)
  - Results weighted for optimal prediction of annual growth (1, 4, 9, 17 based on regression results)

## Results: Average Error in Annual Growth Units

- End-of-year tests had no average bias.
- Simple averages led to underestimates of annual growth.
- Maximum score method led to serious overestimates.
- Matched content and projection scores had little mean bias.

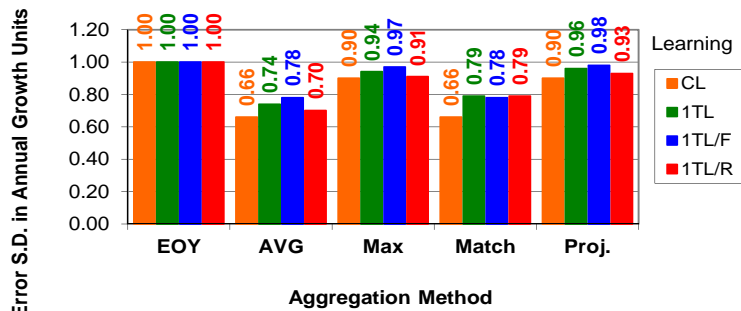
**Avg. Error in Estimating Annual Growth For Each Learning Model**



## Standard Deviation of Estimation Errors

- Standard deviations were reduced for the matched averages approach.
- Slightly reduced for the projection approach

S.D. of Error in Estimating Annual Growth  
For Each Learning Model



## Further Research Needs: A Partial List

- Research on Assessment Design
  - Research on Test Content
    - Content expert judgments of how and when content covered in each grade is taught
  - Research on Learning Models and Aggregation Methods
    - Using pilot versions of actual assessments
- Research on Assessment Use and Impact
  - Research on How Through-Course Assessment Results are Used in Practice
  - Research on Impact of Through-Course Assessments
    - On instruction
    - On subsequent student learning

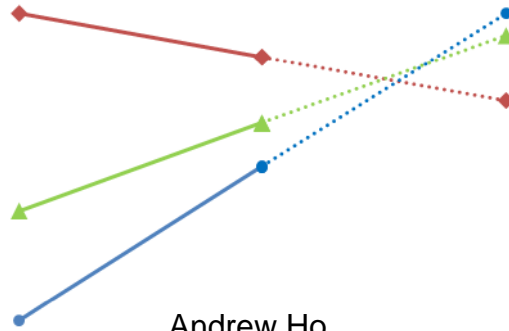
## *Recommendations*

1. Be very cautious in promoting or supporting uses of individual student results!
2. Methods used for aggregating results to estimate end-of-year growth should be based on proven models of how students learn.
3. An end-of-unit testing model is appropriate if test content is matched to material learned in the corresponding quarter.
4. A projection model is appropriate when student learning is essentially continuous throughout the year.

## *Recommendations (Continued)*

5. Short-term research is needed to monitor the different ways, some possibly unintended, that through-course assessment results are used.
6. Longer-term research is needed to gauge the impact of through-course assessment on instruction and improved student learning.

## Supporting Growth Interpretations Using Through-Course Assessments



Andrew Ho

*Harvard Graduate School of Education*  
Webinar for Assessment Professionals  
on Through-Course Summative Assessments  
Thursday, March 10, 2011

## Take-Home Messages

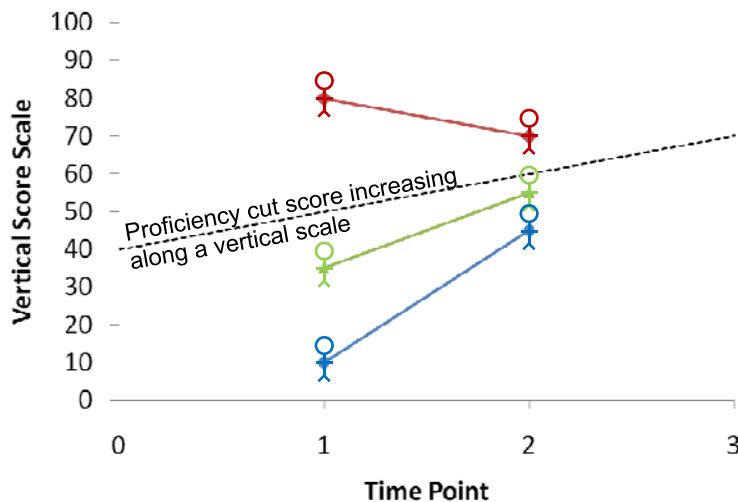
- Recent accountability metrics double as a cautionary tale of unintended consequences.
- Growth holds much promise but is open to stark and surprising contrasts in interpretation.
- This presentation is similar to Laress' in overviewing growth models but asks, in addition, what might happen if we place high stakes on these metrics?
- In light of these tradeoffs, the consortia might consider being realistic about their promise of through-course formative feedback or scaling back their promise for defensible prediction of career and college readiness.

2

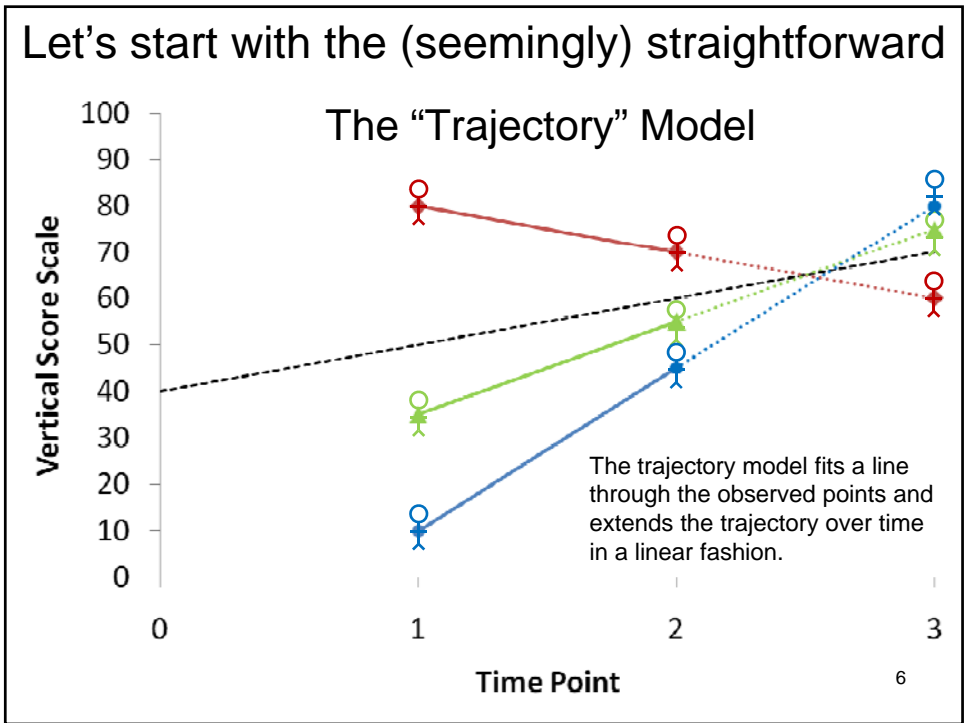
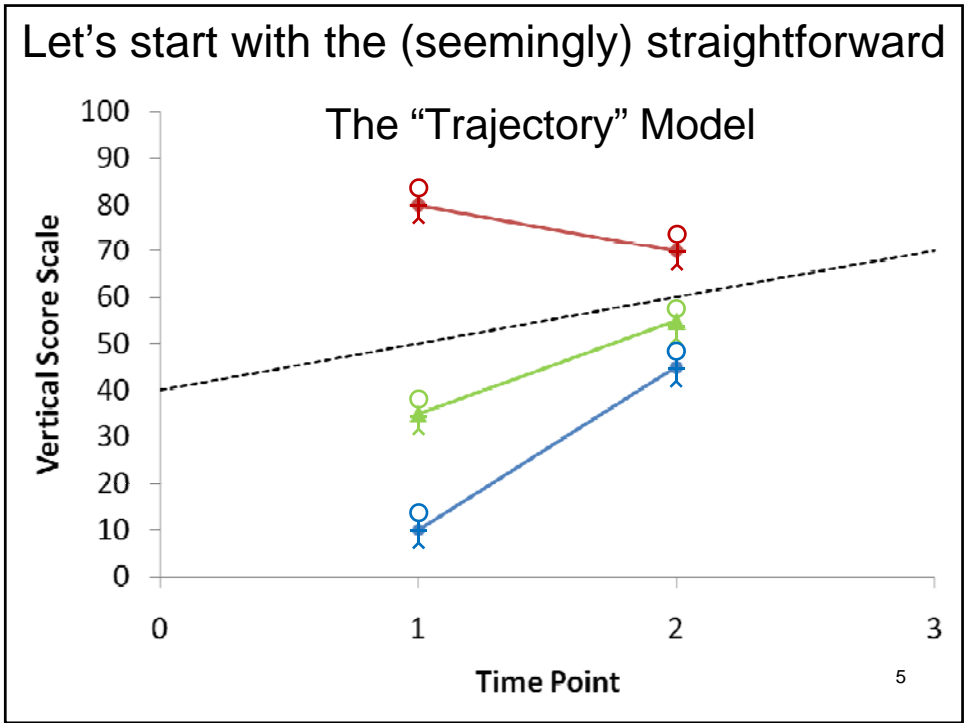
### Through-Course Summative Assessments (TCSAs): A New Hope?

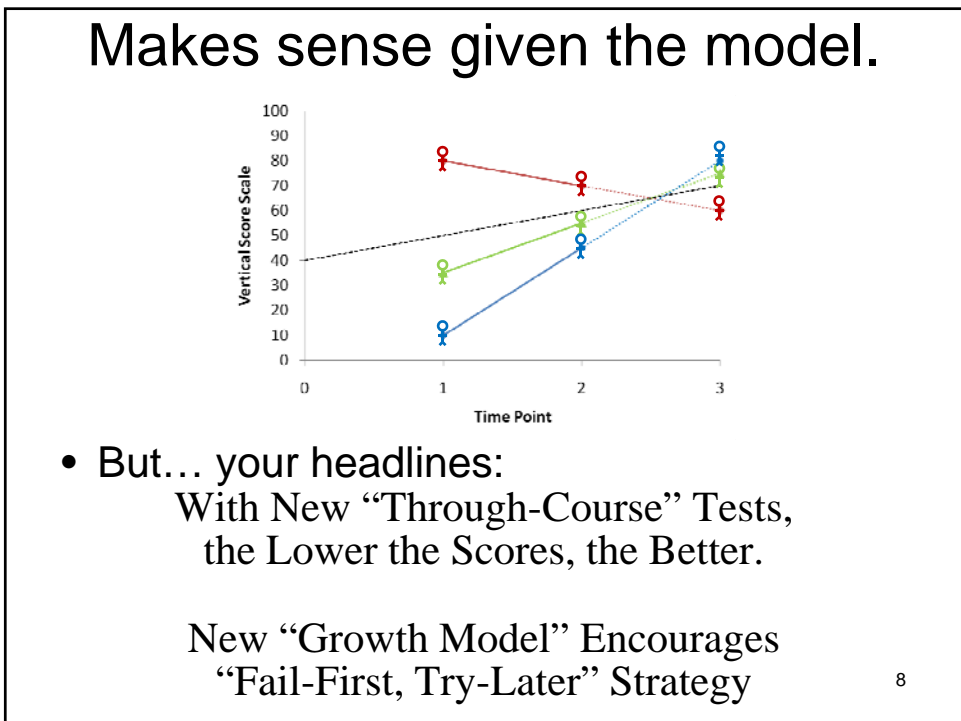
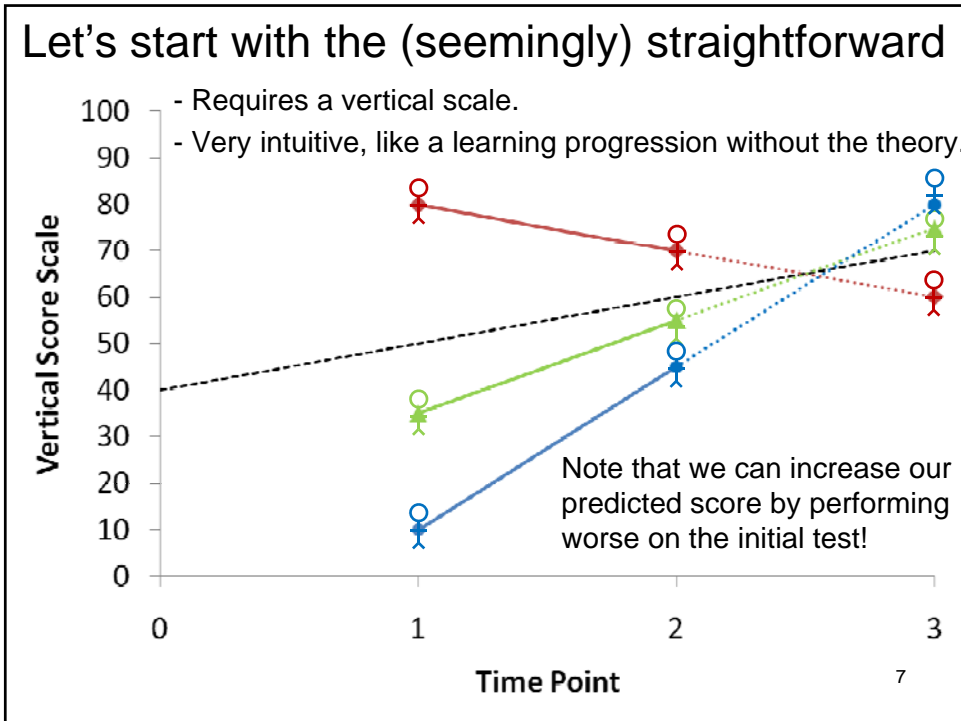
- Multiple formative assessments to support a) instruction and b) defensible school-level (and teacher-level?) classifications and decisions in an accountability model.
- The SBAC and PARCC proposals rely heavily on growth and progress to build an interpretive argument for a) their model for supporting and informing instruction and b) their model for supporting accountability decisions.
  - For the latter, a particular emphasis on growth and progress towards career and college readiness. <sup>3</sup>

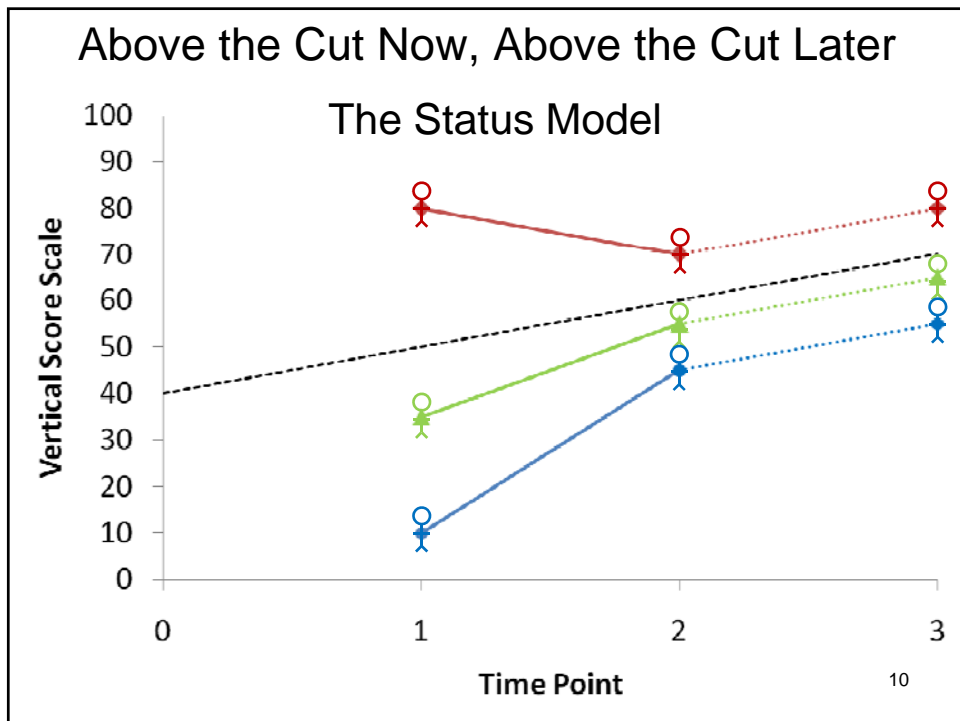
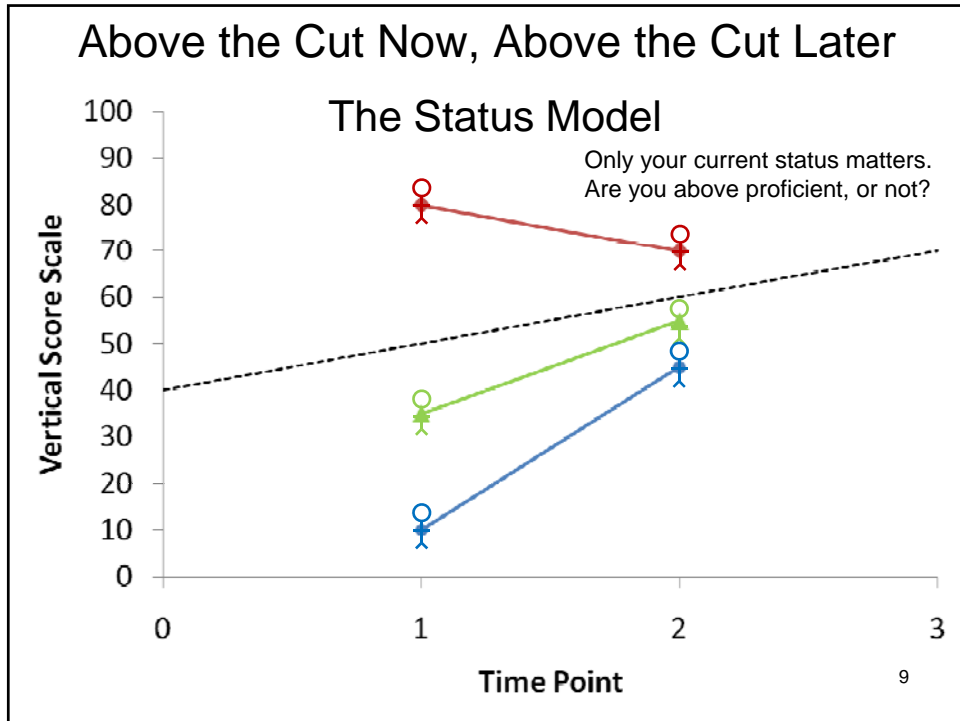
### A Hook: Which Student is “On Track”?



- There are stark contrasts and tradeoffs between classification approaches, accuracy, and gameability. <sup>4</sup>







## The Status Model, Implications

- The Time 2 score has a weight of 1. There is no other variable that matters.
- We would like to maximize our Time 2 score, but our Time 1 score is not a factor.
- Your headlines:

With New “Through-Course” Tests,  
More Testing, Less Accountability.

New “Growth Model” Doesn’t Measure Growth

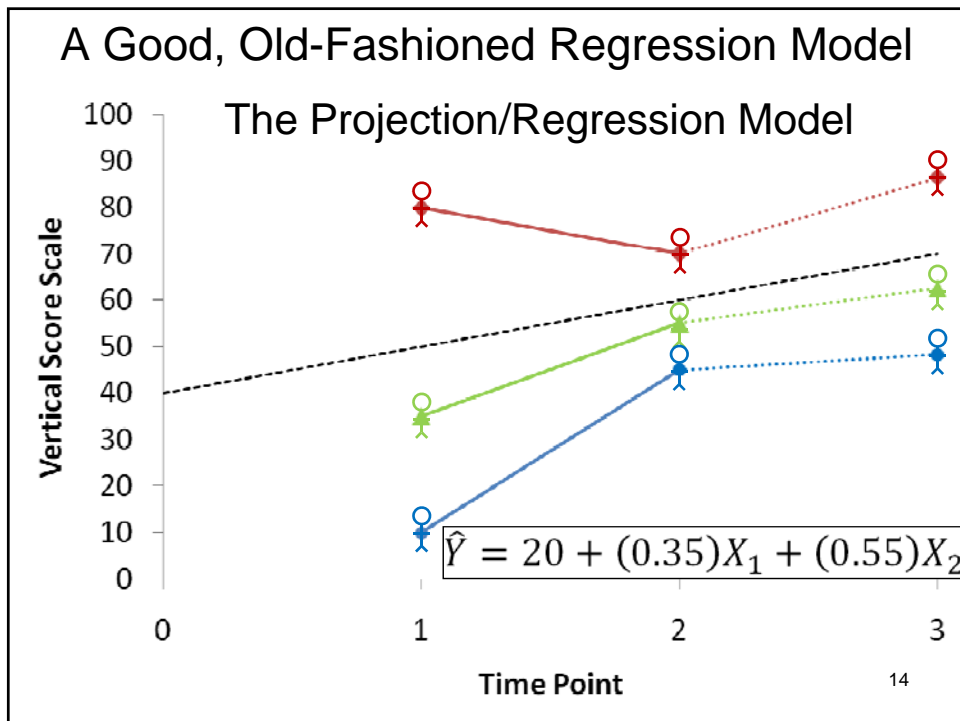
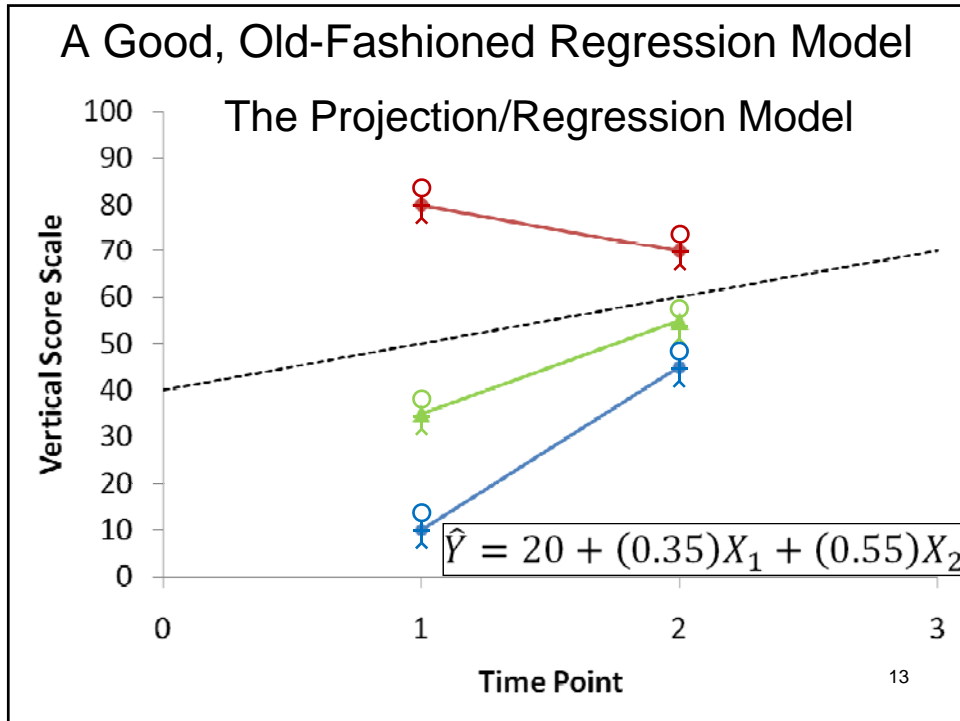
11

## The Projection/Regression Model

- Let’s use data from our outcome test,  $Y$ , and estimate the best weights for  $X_1$  and  $X_2$  using a regression model.
- Note that no current students have any  $Y$  scores, as that test is the prediction we desire for their future.
- A data-driven approach that requires a previous cohort of students with representative data for  $X_1$ ,  $X_2$ , and  $Y$ .

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

12



## The Regression Model, Implications

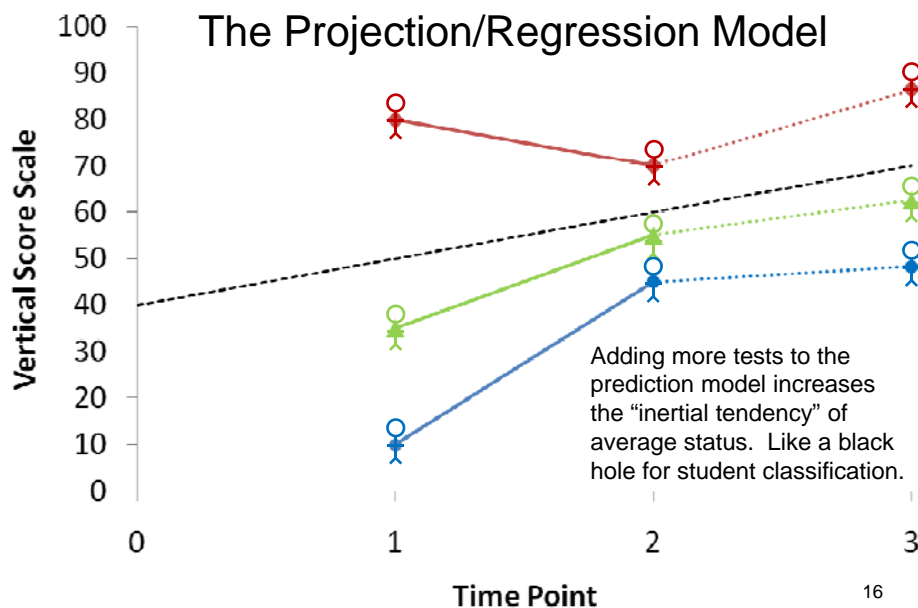
$$\hat{Y} = 20 + (0.35)X_1 + (0.55)X_2$$

- Recent administrations will have higher weights than past administrations.
- Compared to the status model, more stringent for lower scoring students, less stringent for higher scoring students.
- Your headlines:  
 With New “Through-Course” Tests,  
 Initially High-Scoring Students Given Free Pass.

New “Growth Model” Labels Low-Scoring Students Early, Permanently.

15

## A Good, Old-Fashioned Regression Model



16

## Contrasting the Three Models (1)

Trajectory:

$$\hat{Y} = 2X_2 - X_1$$

Status:

$$\hat{Y} = (X_2 - \bar{X}_2) + \bar{Y}$$

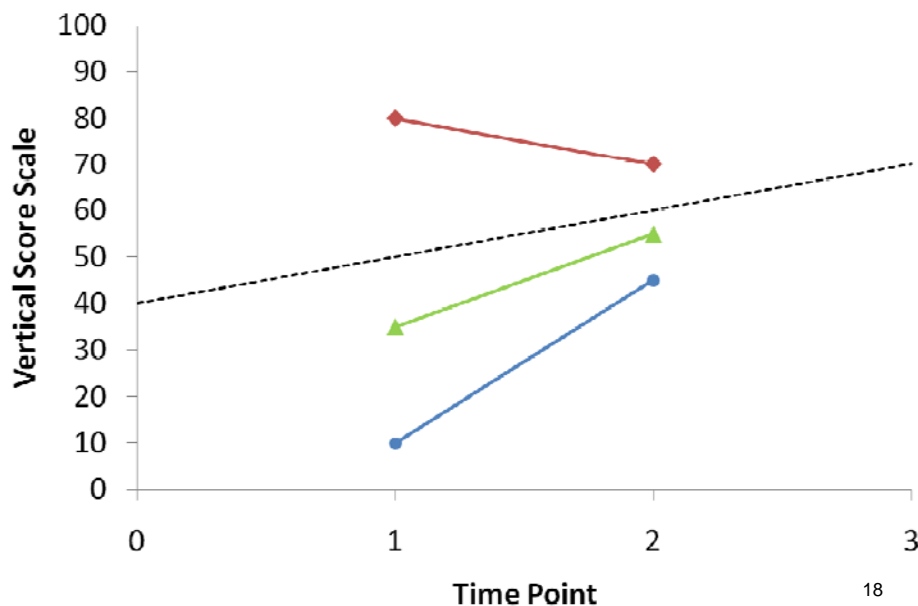
Projection/Regression:

$$\hat{Y} = 20 + (0.35)X_1 + (0.55)X_2$$

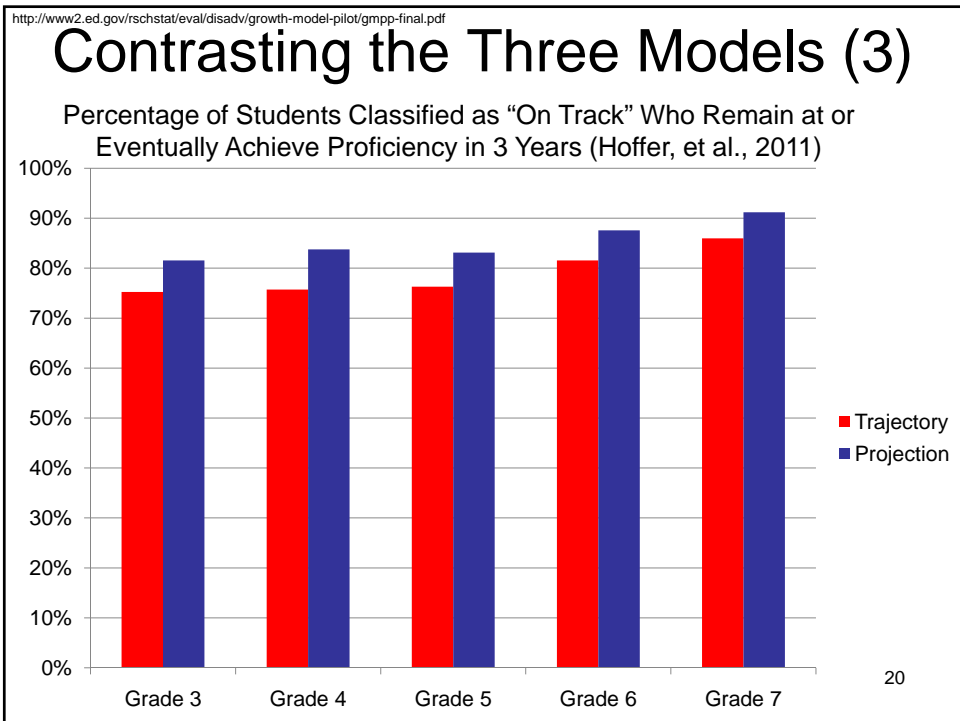
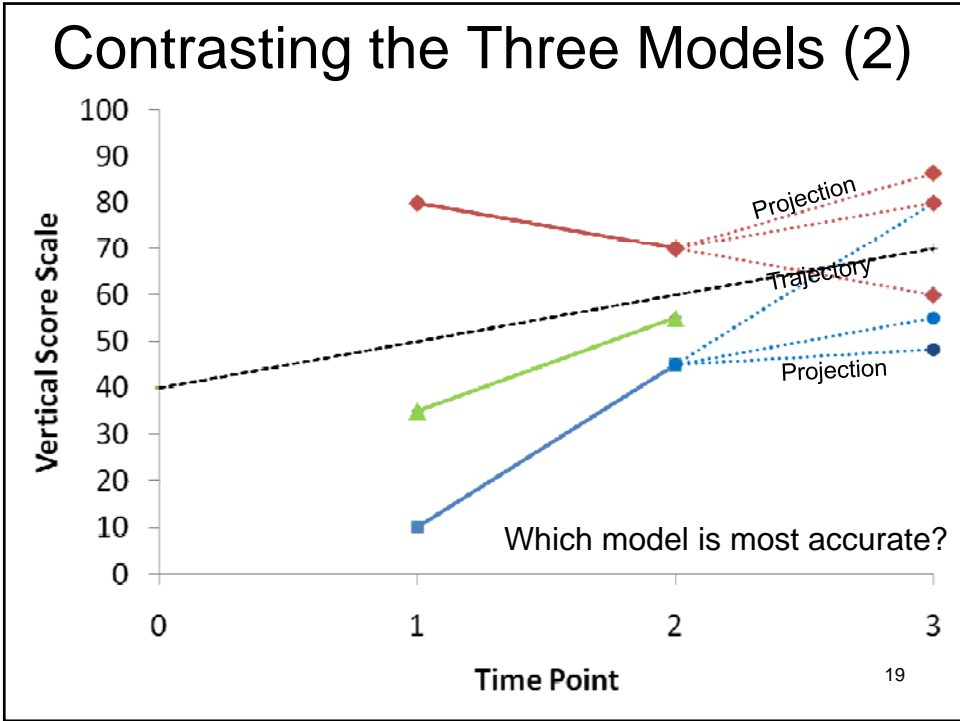
- The clearest algebraic contrast is in the weighting of the initial score,  $X_1$ 
  - Trajectory Model: Negative Weight
  - Status Model: Zero Weight
  - Projection/Regression Model: Positive Weight (but less than the weight for  $X_2$ ).

17

## Contrasting the Three Models (2)



18



## Questions and Answers (1)

- Which model is the most intuitive?
  - The trajectory model, following a momentum metaphor and a clear progression.
- Which model is the most accurate at predicting future outcomes?
  - The projection/regression model. Regression does what regression does.
- Which model is the most game-able?
  - The trajectory model. Generally easier to artificially lower scores than artificially raise them.
- Which model is a growth model?
  - The trajectory model. The regression model is a prediction model that is blind to the order of scores.

21

## Questions and Answers (2)

- Which model is the most costly?
  - Likely to be the trajectory model, as it requires a vertical scale.
- Which model is the most substantively defensible?
  - Likely to be the trajectory model, as it requires a vertical scale. Regression models are atheoretical in that they don't care what variables they are using for the prediction of future outcomes.
- Which model is the best aligned with Title I: Improving the achievement of the disadvantaged?
  - Probably a hybrid, with lower but positive weights on earlier scores, sacrificing both interpretability and predictive utility.

22

## Take-Home Messages

- Recent accountability metrics double as a cautionary tale of unintended consequences.
- Growth holds much promise but is open to stark and surprising contrasts in interpretation.
- This presentation is similar to Laress' in overviewing growth models but asks, in addition, what might happen if we place high stakes on these metrics?
- In light of these tradeoffs, the consortia might consider being realistic about their promise of through-course formative feedback or scaling back their promise for defensible prediction of career and college readiness.

23