

# Interim Assessments & Teacher Evaluation

**NCSL Legislative Summit  
San Antonio, TX  
August 9, 2011**

Anthony Milanowski

Westat

With Contributions from Allan Odden

University of Wisconsin-Madison

# Overview of Presentation

1. How Teacher Evaluation Has Changed
2. Is Policy Ahead of Research?
3. Teacher Effectiveness: A Multiple Measures Approach
4. Advantages Adding of Interim Assessments to the Mix

# How Teacher Evaluation Has Changed!

- Prior to late 90's: no differentiation, no credibility, little consequential use
- New Teacher Project: “The Widget Effect”
  - Almost all teachers rated as “satisfactory” or above → old paradigm is broken
- Measuring Effective Teaching project of the Gates Foundation:
  - Value added measures of teacher impact on student achievement on two sets of tests
  - Several teacher rubrics, with video tool to replace direct observations
  - Student survey

# How Teacher Evaluation Has Changed!

- NCLB, RttT, and TIF are all focusing on teacher effectiveness
  - Requires new measures: new teacher evaluation systems that include more rigorous practice assessment with a measure of effectiveness based on student outcomes
  - Scores of states and districts working on this issue
- This will likely be part of ESEA reauthorization
- *The question is not whether teacher evaluation will change but how it will be changed*

# The New Paradigm

- More Rigorous Evaluation of Practice Using Multi-Level Rating Scales
- Inclusion of Estimated Impacts on Student Achievement or Growth
- Standard Prescription: Multiple measures consisting of instructional practice measure (e.g., teacher evaluation ratings) + Gain, growth, or value-added based on state standards-based assessments

# The New Paradigm

- Consequential Use of New Evaluation Measures:
  - a) For tenure
  - b) For tracking equitable distribution of effective teachers
  - c) For distributing and placing effective teachers
  - d) For dismissing ineffective teachers
  - e) For compensating teachers

# Rating Scale/Rubric Example (Cincinnati Public Schools)

	Distinguished	Proficient	Basic	Unsatisfactory
<b>Teacher Interaction with Students</b>	Teacher interactions with all students demonstrate a positive, caring rapport and mutual respect. Interactions are inclusive and appropriate.	Teacher interactions with all students demonstrate respect. Interactions are inclusive and appropriate	Teacher interactions with students are generally appropriate.	Teacher interactions with students are negative, demeaning, and/or inappropriate.

# Is Policy Ahead of Research (and Reality)?

1. No agreement yet on the “right” value-added or growth model
2. Incomplete teacher-student linkage
3. Estimates based on one year of value-added are not stable
4. Many assessments are not instructionally sensitive nor instructionally useful (results come too late to tell teachers how to help their students and how to improve instructional practice)
5. Up to 70% of teachers do not teach tested subjects

# A True Multiple Measures Approach Would:

- Use at least two years of a value-added or growth estimates if consequential decisions are to be made.
- Assess student achievement using more than one test
- Require that measurement results converge when making tenure and dismissal decisions

# Converging Evidence

	Student Outcome Rating			
Teaching Practice Rating	1	2	3	4
4 = Advanced	?	?	Reward	Reward
3 = Proficient	?	On Notice	Tenure	Reward
2 = Basic	Dismiss	On Notice	On Notice	?
1 = Unsatisfactory	Dismiss	Dismiss	?	?

# Advantages of Adding Short Cycle Assessments to the Mix

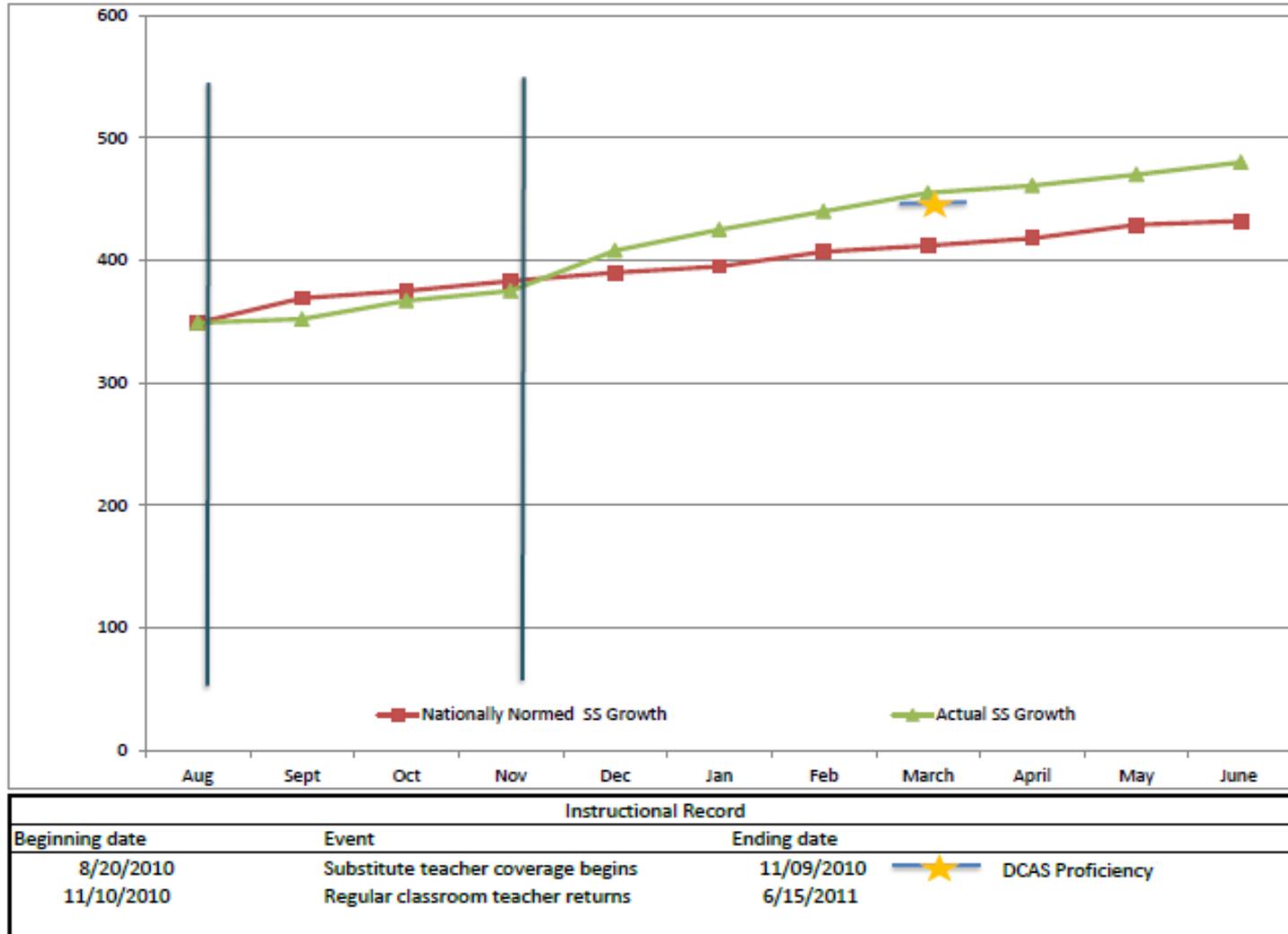
- Few states and districts actually have viable multiple indicators; in most cases only indicator a “value added” measure derived from state summative, accountability tests
- Most teachers do not like value-added measures using end of year state summative tests; don’t understand them; don’t like state tests
- But many use and value interim or short cycle assessments

# Short Cycle Assessments

- Are given several times during the year
- Administered in computer-based format so provide immediate feedback to teachers for use in instructional improvement and change (STAR is one example)
- Vertically aligned scales so can compare scores to show growth across months and years
- Interim assessments often cover more grades than just the NCLB 3-8 + 1 HS year

# Short Cycle Assessment Growth

Evidence of Student Learning - Mr. Taylor, Grade 5



# Tracking Growth

- Interim assessments given monthly
- Student data aggregated to classroom
- Red squares are progress line for similar classes of students in a state (or nation)
- Green triangles are actual class progress
- Yellow star is state proficiency level
- Shows growth during the months of just the academic year

# Information Provided By Tracking Chart

- Modest student learning when the class had a substitute teacher
- Growth happened when regular teacher returned
- Actual class growth (green triangles) was much greater than the reference norm (red squares)
- In value added terms, the class would have a high value added – performance growth was above the average (the red square trend line) for this typical classroom

# Adding Short-cycle Assessments

## Benefits Teachers

- Because such assessments are frequent, teachers get feedback that they can use to adjust instruction *before* the state test
  - Could reduce stress by giving teachers more tools for preparing their students for the state test
- Teachers can see if student achievement is improving, and if assessments are linked to state proficiency levels, whether students are on track to proficiency
- Provides another measure of performance that compliments value added or growth on state test; reduces weight placed on that test in judging teaching effectiveness

# Report Card Model

Performance Domain	Performance Dimensions	Score Levels	Requirement for Being Considered Effective
Instructional Practice	Planning & Assessment Classroom Climate Instruction	1-4 1-4 1-4	Rating of 3 or higher on all dimensions
Professionalism	Cooperation Attendance Development	1-4 1-4 1-4	Rating of 3 or higher on all dimensions
Student Growth, Gain, or VA on State Assessments	Math Reading/ELA Other Tested Subjects	Percentiles in state/district distribution for each subject	Being in the 3 <sup>rd</sup> Quintile or Higher for All Tested Subjects
Student Growth on Short Cycle Assessment	Math Reading	Avg. Growth Curve Translated into Predicted State Test Scale Score Change	Predicted Gain Over Year Sufficient to Bring Student from Middle of “Basic” Range to “Proficient”

# For Techies

## Short Cycle Assessments Can Improve Measurement of Teacher Effectiveness:

- More data points allow estimation of a growth curve
- The growth curve represents learning within a single school year; no summer to confuse attribution
- The slope of the average growth curve or average difference between predicted end points provides another indicator of teaching effectiveness
- Combining with growth, gain, or value-added based on state assessments provides multiple measures of productivity
- If linked to state assessments, can predict school year proficiency growth

**For a copy of this presentation  
email:**

[Anthony\\_Milanowski@westat.com](mailto:Anthony_Milanowski@westat.com)

See also:

[http://www.wcer.wisc.edu/publications/workingPapers/Working\\_Paper\\_No\\_2011\\_02.pdf](http://www.wcer.wisc.edu/publications/workingPapers/Working_Paper_No_2011_02.pdf)