Assessing the quality of teachers and schools: What we have learned, and where we are headed

Cassandra Guarino, PhD
Professor of Public Policy and Education
UC Riverside

Presented to UC Sacramento, February 2, 2017

Brief historical outline of push for school and teacher effectiveness measures

- 2002: School accountability begins under GW Bush administration with reauthorization of ESEA called **No Child Left Behind** and establishment of widespread achievement testing
- 2009-2015: Teacher accountability takes off under Obama administration with **Race to the Top** and **Flexibility Waivers**
- Late 2015: New reauthorization of ESEA under Obama is called **Every Student Succeeds Act**, which eases off teacher accountability but retains school accountability

NCLB, 2002 bipartisan effort, required school evaluation

Coauthors:
John Boehner (R-OH)
George Miller (D-CA)
Ted Kennedy (D-MA)
Judd Gregg (R-NH)

- Stated purpose: "To close the achievement gap with accountability, flexibility, and choice, so that no child is left behind"
- Ushered in an era of accountability fraught with change and controversy

Main provisions of NCLB

- Schools were required to bring all students to proficiency by 2014
- In the interim, they were required to make "Adequate Yearly Progress" (AYP)
- States were given the flexibility to determine what constituted AYP, within certain requirements

AYP requirements

- To make AYP, schools had to satisfy requirements for subgroups of students as well as overall
  - Socioeconomically disadvantaged
  - Students with disabilities
  - English learners
- Consequences for schools that did not make AYP
  - 2 consecutive years: identify for improvement
  - 3 consecutive years: free tutoring and supplemental services must be offered
  - 4 consecutive years: corrective action, such as replacing staff
  - 5 consecutive years: restructuring plan (e.g., closure, conversion to charter, state takeover, management organization)
  - 6 consecutive years: implement plan

AYP calculation in California

- Typically incorporated:
  - Annual Measurable Objective proficiency on standardized tests
  - Academic Performance Index based on proficiency and growth
  - Graduation rates for high schools only

Source: http://www.cde.ca.gov/ta/ac/ay/documents/aypinfoguide13.pdf

Many schools failed to make AYP

- Center on Education Policy report estimates for 2011:
  - 48% of schools in US did not make AYP
  - 66% of schools in CA did not make AYP
- States had a lot of control over how stringent they wanted to make their own tests and standards

Source: file:///C:/Users/cguarino/Downloads/Usher_Report_AYP2010-2011_110112.pdf

Federal government also pushed teacher evaluation under Obama

- Arne Duncan, Obama's longest serving Secretary of ED, shifted the theory of action from the school to the teacher
- Rationale: Effective teachers were key to student learning
- Schools and districts needed to measure how effective teachers were
- Duncan had two unprecedented vehicles to promote his reform:
  1) Race to the Top
  2) Flexibility Waivers

Race to the Top and Flexibility

- Duncan instituted a contest called Race to the Top in 2009, using ARRA (stimulus) discretionary funding of more than $4 billion
- NCLB's impossible 2014 provision allowed Duncan to grant Flexibility Waivers to all states that did not meet full proficiency by 2014
- To win the contest or to gain a waiver, states had to show they had a plan for improving instruction that included:
  - … evaluating and supporting teachers and principals
  - … evaluations had a student test score component
- In this way, Duncan promoted widespread change in teacher evaluation

How can we assess effective teaching?

- Two main methods
- Look at the outcomes of a teachers' students
  - Estimate a teacher's contribution to student test scores using a statistical "value-added" model
- Observe a teacher in action
  - Form a subjective opinion (as in traditional principals' or peer evaluations)
  - Use some sort of "quality" rating rubric
- Most teacher evaluation systems combine both types of measures, although some have no value-added component

What is value-added?

- Idea is to determine the value that a teacher has added over the course of the year
- Value is computed in terms of improvements in student achievement
- Value-added always takes prior achievement into account

Simple example-just average scores

Mrs. Jones's class			Ms. Smith's class	
Students	End of year test scores		Students	End of year test scores
Sally	150		Josh	220
Billy	270		Amy	370
S3	330		S3	410
S4	100		S4	350
S5	125		S5	490
S6	158		S6	215
S7	290		S7	310
S8	210		S8	375
S9	300		S9	280
S10	250		S10	330
Average	218.3		Average	335

Which teacher has higher achieving students?

Now factor in prior test scores

Mrs. Jones's class				Ms. Smith's class			
Students	End of year test scores	Prior year test scores	Test score gain	Students	End of year test scores	Prior year test scores	Test score gain
Sally	150	100	50	Josh	220	202	18
Billy	270	220	50	Amy	370	325	45
S3	330	270	60	S3	410	380	30
S4	100	89	11	S4	350	300	50
S5	125	75	50	S5	490	470	20
S6	158	90	68	S6	215	180	35
S7	290	230	60	S7	310	275	35
S8	210	175	35	S8	375	325	50
S9	300	280	20	S9	280	235	45
S10	250	205	45	S10	330	290	40
Average	218.3	173.4	44.9	Average	335	298.2	36.8

Which teacher "adds more value"?

But are student learning gains all due to their teacher?

- What other factors might contribute to this growth?
  - Families
  - Tutors
  - Peers
  - Differences in students' own efforts and abilities
  - Differences in student ability (or disabilities)
  - Neighborhood environments
- Some of the better value-added models can adjust for some of these factors
- But concern still remains that they may not accurately measure a teacher's effectiveness

Applications of value-added have sparked controversy

- Point of contention in the 2012 Chicago teachers' strike
- Public releases have angered teachers and unions
  - Los Angeles Times has published measures for LAUSD teachers in grades three to five on the web since 2010
    - http://projects.latimes.com/value-added/
  - New York City court battles, WSJ published measures for grades four through eight after 2012
    - http://online.wsj.com/articles/SB10001424052970203918304577243282490252956

LA Times example

Los Angeles Teacher Ratings

Recommend 0 Tweet 0

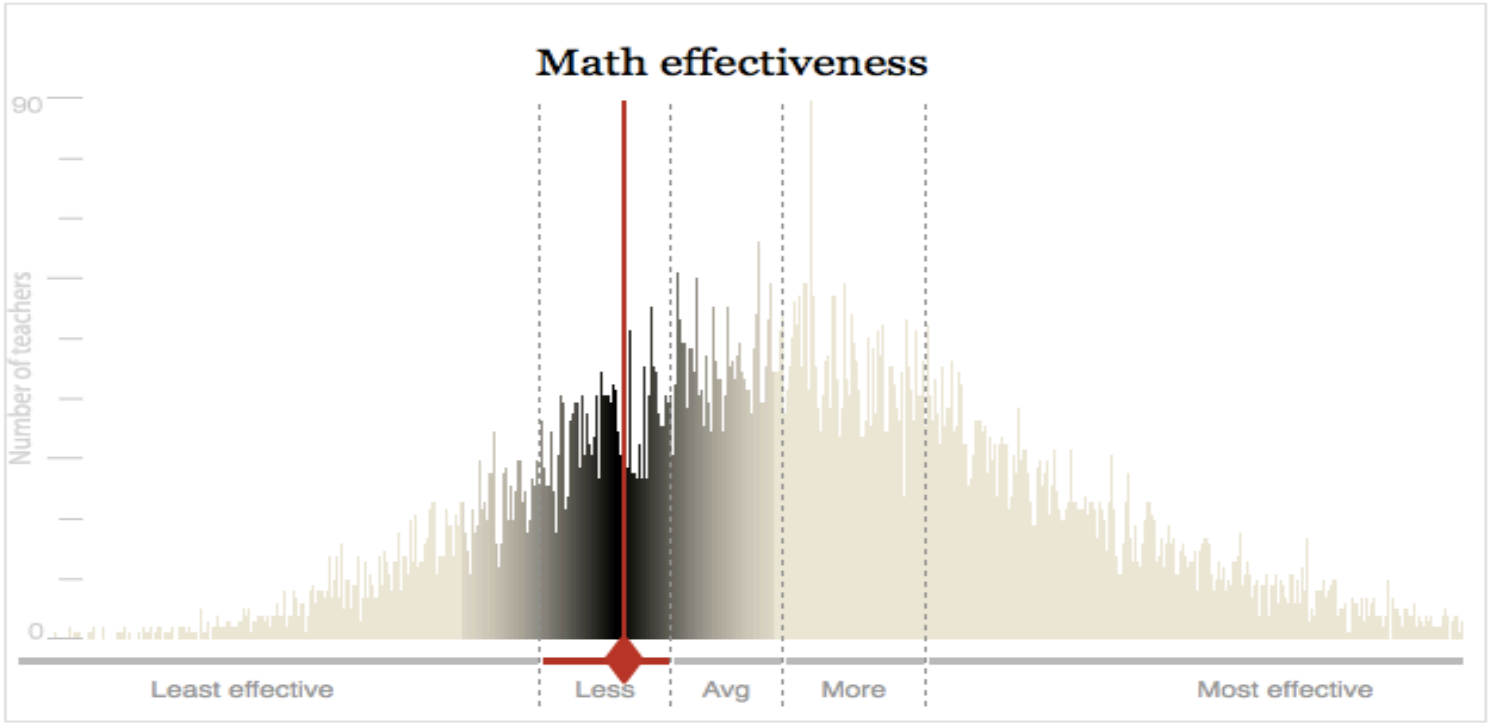
See 2009 rating

Claudia G. Martinez

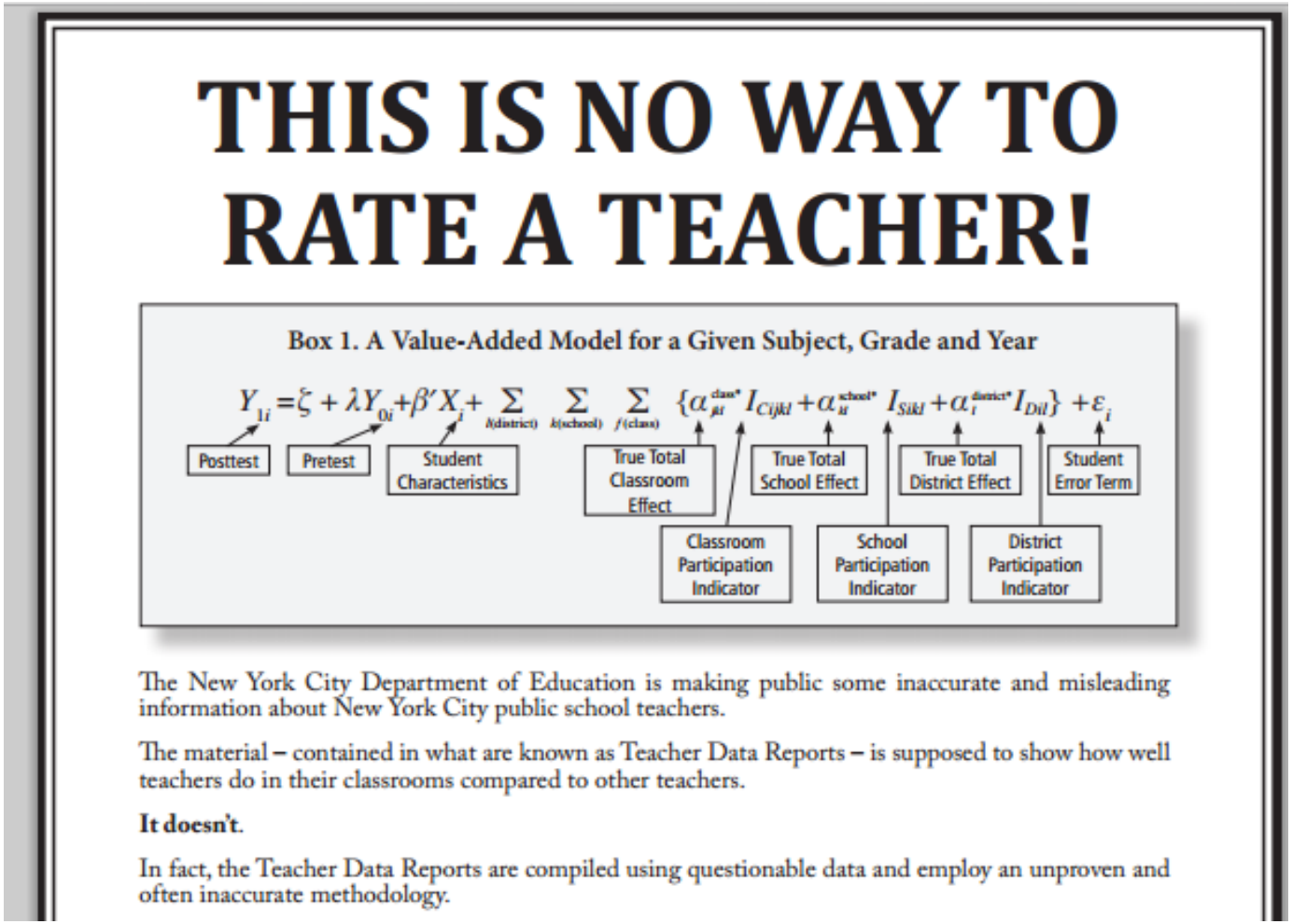
A 4th grade teacher at Meyler Street Elementary in 2010

These graphs show a teacher's "value-added" rating based on his or her students' progress on the California Standards Tests in math and English. The Times' analysis used all valid student scores available for this teacher from the 2003-04 through 2009-10 academic years. The value-added scores reflect a teacher's effectiveness at raising standardized test scores and, as such, capture only one aspect of a teacher's work.

← Drag graph for more →



+ Union backlash over public release



THIS IS NO WAY TO RATE A TEACHER!

Box 1. A Value-Added Model for a Given Subject, Grade and Year

$$Y_{1i} = \xi + \lambda Y_{0i} + \beta' X_i + \sum_{k(\text{district})} \sum_{k(\text{school})} \sum_{j(\text{class})} \{\alpha_{\mu t}^{\text{class}} I_{Cijkl} + \alpha_{u}^{\text{school}} I_{Sikl} + \alpha_{t}^{\text{district}} I_{Dil}\} + \varepsilon_i$$

Posttest; Pretest; Student Characteristics; True Total Classroom Effect; True Total School Effect; True Total District Effect; Student Error Term; Classroom Participation Indicator; School Participation Indicator; District Participation Indicator

The New York City Department of Education is making public some inaccurate and misleading information about New York City public school teachers.

The material – contained in what are known as Teacher Data Reports – is supposed to show how well teachers do in their classrooms compared to other teachers.

It doesn't.

In fact, the Teacher Data Reports are compiled using questionable data and employ an unproven and often inaccurate methodology.

See full UFT ad at:
file:///C:/Users/guarino/Documents/Teaching%20H200%20Fall%202014/uft-teacher-data-reports-formula-ad%20(1).pdf

Pros and cons of using value-added for teacher evaluation

- Pros
  - Based on outcomes we care about
  - Data driven, therefore objective
  - Better than rewards to teachers based solely on experience and education
  - Relatively economical since standardized tests are ubiquitous due to NCLB
- Cons
  - May be imprecise due to small numbers of students per teacher
  - Inaccurate matching can occur, especially with very mobile students
  - Based on tests that may not measure everything we want students to know
  - Doesn't give teachers much information on how to improve

How good are value-added measures?

- Researchers have tried several approaches to determine the accuracy of value-added measures
- Experimental approaches tell us they send a credible signal, but with some noise
  - Kane & Staiger 2008, Chetty et al. 2013
- Simulation studies show some approaches are clearly better than others
  - Guarino, Reckase & Wooldridge, 2015
- Comparisons of estimates over time show a fair amount of instability
  - Aaronson et al. 2007, McCaffrey et al. 2009, Koedel & Betts 2007
- Instability can be greater for teachers who teach low-performing students
  - Stacy, Guarino & Wooldridge, under review

My research on value-added models

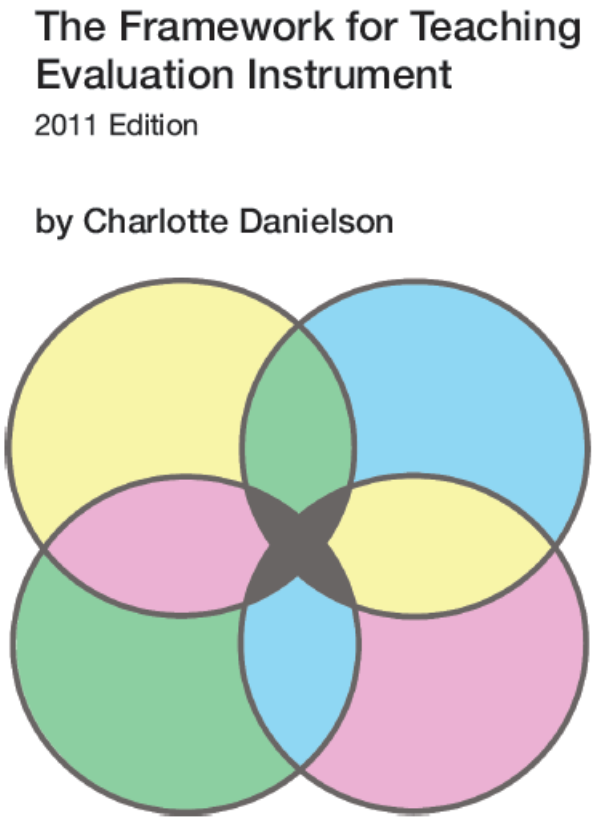
- Assesses the strengths and weaknesses of different value-added methods
- We find that some methods are less biased than others
  - Guarino, Reckase & Wooldridge 2015
  - Dieterle, Guarino, Reckase & Wooldridge 2015
  - Guarino, Reckase, Stacy & Wooldridge 2015
  - Guarino, Maxfield, Reckase, Thompson & Wooldridge 2015
- Surprisingly, the best methods are not necessarily the most widely used
- Weaker, more biased, computations are more popular
- This holds for both teacher and school evaluation

Classroom observations increasingly used as teacher evaluation measures

- Nearly all teacher evaluation systems incorporate some form of classroom observation measure
- In the past, school principals walked into a classroom on occasion to observe a teacher for a few minutes
- For the most part, principals gave teachers good ratings
- Now, there is pressure to be more discriminating and systematic in the approach to classroom observations
- Several observation "rubrics" have been developed to guide the process

Framework for Teaching (FFT)

- FTT is a popular rubric that has formed the basis for many others
- Covers four domains
  - Planning and preparation
  - The classroom environment
  - Instruction
  - Professional Responsibilities
- Complex rubric contained in 104-page document



Scoring teachers with rubrics is challenging

- Measures of Effective Teaching project, funded by the Gates Foundation, studied five different types of observation rubrics
- Despite extensive training, observers varied a great deal in their scores
- Scores on rubrics varied
  - From rater to rater on the same lesson
  - From rubric to rubric on the same lesson
  - From lesson to lesson for the same teacher using the same rubric
- In statistical terms, these ratings have low reliability
- See pg. 35 of "Gathering Feedback for Teaching"
  - http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Pros and cons of using classroom observations to measure teacher performance

- Pros
  - Provides a systematic analytic guide to effective teaching
  - Gives teachers helpful information regarding their teaching
  - Forms basis for professional development
  - Engages teachers and administrators in dialogue over instruction
- Cons
  - Scoring is subject to variability and subjectivity
  - Measure of pedagogical technique may not necessarily measure how much students learn
  - Time-consuming to train and implement, therefore expensive
  - Burdensome for school administrators

Variation in teacher evaluation

- Wide variation across states and districts in the implementation of teacher evaluation
- California makes relatively weak efforts to promote teacher evaluation
  - Very hard to link student test scores to teachers
- Indiana, NYC Public Schools, DCPS, etc., much more assertive
- Even in states and districts that use student test scores to evaluate teachers, they are usually counterbalanced by observation scores
- Very few teachers get rated as ineffective (just like before)

Shifting winds in Washington

- Prior to midterm elections in 2014, the Obama administration softened attitude toward using testing in accountability
- "Too much testing can rob school buildings of joy, and cause unnecessary stress. This issue is a priority for us, and we'll continue to work throughout the fall on efforts to cut back on over-testing."
  - *---Arne Duncan in letter "Listening to Teachers on Testing"*
  - http://blogs.edweek.org/edweek/curriculum/2014/08/us_ed_sec_duncan_slams_excessi.html
- Changed Flexibility Waivers to allow an extra year before using test scores in teacher evaluations
  - http://blogs.edweek.org/edweek/campaign-k-12/2014/08/ed_announcment.html

ESSA, 2015 bipartisan effort

Coauthors:
Patty Murray (D-WA)
Lamar Alexander (R-TN)
John Kline (R-MN)
Bobby Scott (D-VA)

ESSA continued school evaluation but <u>did not include</u> teacher evaluation based on test scores

ESSA added a new component to school evaluation

- Continued testing in reading and math grades 3-8 and once in high school
- States still submit accountability plans to ED
- Can pick their own accountability goals, measured against four indicators, which must include:
  - Proficiency on tests
  - Proficiency in English language
  - Another academic indicator, such as test score growth, broken out by subgroups or graduation rates for high schools
  - Non-academic indicator, such as student engagement, school climate and safety, college readiness
- States can decide how to weight the indicators

In CA, LCFF shift from state to local accountability allows varied approaches

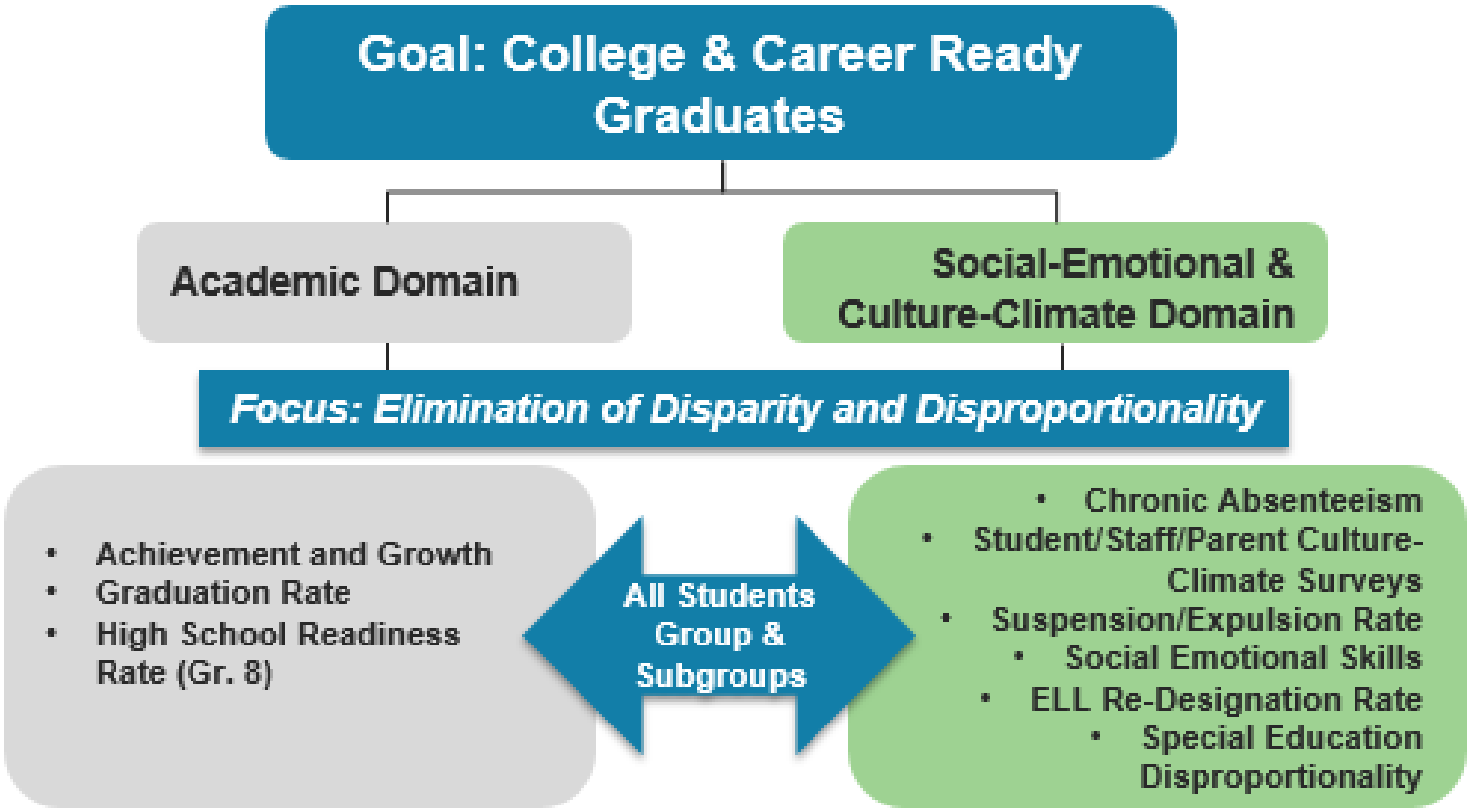
58 counties

1,025 school districts

Slide taken from CORE presentation: http://coredistricts.org/wp-content/uploads/2016/12/CERA-plenary-session-12.1.16-v6.11.30.16.pdf

CORE districts in CA designed School Quality Improvement Index

- CORE districts:
  - Fresno, Garden Grove, Los Angeles, Long Beach, Oakland, Sacramento City, San Francisco, Santa Ana



Source: http://coredistricts.org/wp-content/uploads/2016/12/CERA-plenary-session-12.1.16-v6.11.30.16.pdf

Several issues in school effectiveness remain to be explored

- How do you weight the multiple measures?
- How do you determine cutoffs to categorize effective or ineffective schools?
- Note: new California accountability plan will not provide one single school effectiveness number but a "Dashboard" of multiple indicators similar to those in the CORE model

What have we learned?

- Important to keep policy stakes low until stakeholders become familiar with and see the usefulness of measures
- Important to find ways to include multiple measures of student development, social-emotional and behavioral as well as academic
- Helpful to heed research about pros and cons of methods and which methods are best before implementing them in high stakes policies

END

Thank you!

Cassandra.Guarino@ucr.edu