CENTER FOR EDUCATION POLICY ANALYSIS at STANFORD UNIVERSITY

cepa

How Stable Are Value-added Estimates Across Years, Subjects And Student Groups?

Susanna Loeb
Christopher A. Candelaria
Stanford University

"Value-Added" Models

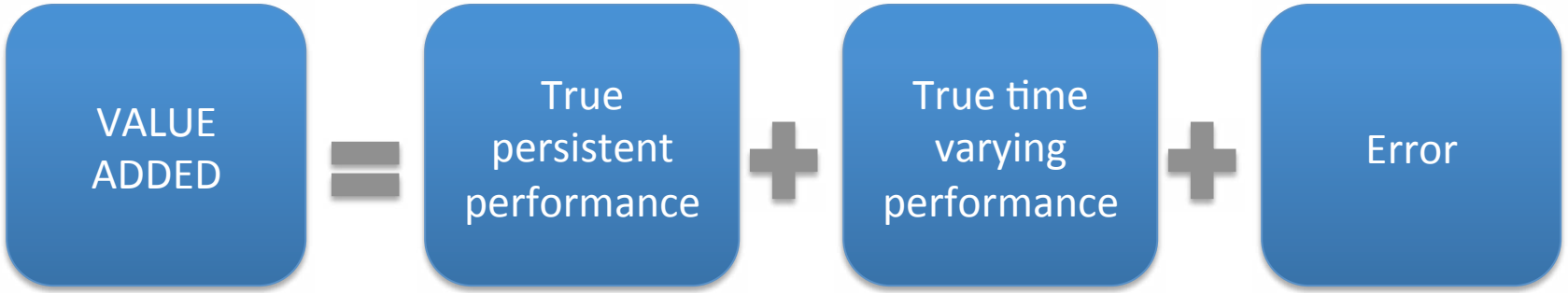
- Value-added models
  - Measure teacher performance by the test score gains of their students
  - Adjust for a variety of factors such as students' prior academic achievement, student background characteristics, and classroom characteristics
- Advantages
  - Measure effectiveness based on student outcome
- Potential Disadvantages
  - Validity
    - Is the underlying test a measure of outcomes that we care about?
  - Bias
    - Do we appropriately attribute value-added to teachers?
  - Reliability
    - Are the estimates precise or is there a lot of error?

Stability

- Measures vary
  - Year to year
  - Subject to subject
  - Student group to student group
- Source of Variation
  - True differences in a teacher's performance
  - Error
    - Tests are imperfect measures of students' skills and knowledge
    - Idiosyncratic forces affect classrooms each year
  - Difficult to separate error from true differences

Stability Across Time

- Some teachers are better than others at improving standardized test scores
- Teachers have good years and bad years
- Teachers can improve over time (not all instability is bad)
- Tests have errors which are enhanced when measuring gains



Stability Across Years

- **Method I: Correlation**
  - **1.0** if a teacher's value-added measure in one year is perfectly predictive of her score in the following year
  - **0.0** if her score one year tells us nothing about how she will fare the next
  - Range of estimates
    - Study of elementary school teachers
      - 0.6 to 0.8 for math
      - 0.5 to 0.7 for reading
    - Some other studies find lower correlations
    - Similar to other occupations
  - Averaging across years increases ability to predict future years
    - 0.4 correlation of one year data to the next
    - 0.6 correlation of the average of two years data to the next
    - More than two years of data adds a little but not a lot.
- **Method II: Transition Matrixes**
  - Describes the transition of teachers from one part of the distribution in one year to another part in the following year

Table 1

Stability of Teacher Value-Added Across Time: (Percent of Teachers by Row)

	Ranking in year 2 (Quintiles)				
Ranking year 1 (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)
Q1 (Bottom)	43%	29%	14%	10%	4%
Q2	26%	21%	25%	18%	9%
Q3	12%	21%	28%	25%	15%
Q4	10%	19%	19%	28%	23%
Q5 (Top)	8%	11%	11%	19%	50%

Notes: Total teachers = 941.Source: Koedel & Betts (2007)

Table 1

Stability of Teacher Value-Added Across Time: (Percent of Teachers by Row)

	Ranking in year 2 (Quintiles)				
Ranking year 1 (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)
Q1 (Bottom)	43%	29%	14%	10%	4%
Q2	26%	21%	25%	18%	9%
Q3	12%	21%	28%	25%	15%
Q4	10%	19%	19%	28%	23%
Q5 (Top)	8%	11%	11%	19%	50%

Notes: Total teachers = 941.Source: Koedel & Betts (2007)

Table 1

Stability of Teacher Value-Added Across Time: (Percent of Teachers by Row)

	Ranking in year 2 (Quintiles)				
Ranking year 1 (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)
Q1 (Bottom)	43%	29%	14%	10%	4%
Q2	26%	21%	25%	18%	9%
Q3	12%	21%	28%	25%	15%
Q4	10%	19%	19%	28%	23%
Q5 (Top)	8%	11%	11%	19%	50%

Notes: Total teachers = 941.Source: Koedel & Betts (2007)

Stability Across Subjects

- Similar combination of true differences and error
- Most evidence on differences between Math and English language arts for elementary school
- Correlation coefficients in the range of 0.2 to 0.6

Table 2

Stability of Teacher Value-Added Across Subjects(Percent of Teachers by Row)

	Ranking in reading (Quintiles)					
Ranking in math (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	46%	27%	14.%	9%	4%	704
Q2	23%	28%	22.%	17%	9%	622
Q3	16%	24%	23%	21%	16%	580
Q4	8%	18%	25%	28%	21%	612
Q5 (Top)	4%	10%	16%	25%	46%	589
Total	641	671	616	608	571	3,107

Source: Loeb, Kalogrides, & Béteille (Forthcoming).

Table 2

Stability of Teacher Value-Added Across Subjects(Percent of Teachers by Row)

	Ranking in reading (Quintiles)					
Ranking in math (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	46%	27%	14.%	9%	4%	704
Q2	23%	28%	22.%	17%	9%	622
Q3	16%	24%	23%	21%	16%	580
Q4	8%	18%	25%	28%	21%	612
Q5 (Top)	4%	10%	16%	25%	46%	589
Total	641	671	616	608	571	3,107

Source: Loeb, Kalogrides, & Béteille (Forthcoming).

Table 2

Stability of Teacher Value-Added Across Subjects(Percent of Teachers by Row)

	Ranking in reading (Quintiles)					
Ranking in math (Quintiles)	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	46%	27%	14.%	9%	4%	704
Q2	23%	28%	22.%	17%	9%	622
Q3	16%	24%	23%	21%	16%	580
Q4	8%	18%	25%	28%	21%	612
Q5 (Top)	4%	10%	16%	25%	46%	589
Total	641	671	616	608	571	3,107

Source: Loeb, Kalogrides, & Béteille (Forthcoming).

Stability across Student Populations

- Substantial variation among student groups
  - Demographic, academic, and otherwise
  - Few studies examine the extent to which teachers are more effective with one group of students than with another
- Variation by prior student achievement
  - Some variation, but largely similar
  - Correlation of 0.4 in one small study
- Variation by ELL status
  - One study: 0.4-0.6 math, 0.3-0.4 reading (compared with 0.7 and 0.6 for randomly separated group)

Table 3

Stability of Teacher Value-Added Between English Learners and Others (Math)

	Ranking for Non-English Learners (Quintiles)					
Ranking for English Learners	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	50%	25%	15%	7 %	4%	467
Q2	23%	32%	23%	14%	5%	480
Q3	15%	23%	28%	21%	12%	484
Q4	9%	15%	25%	30%	20%	480
Q5 (Top)	4%	4 %	10%	28%	59%	475
Total	409	525	541	504	407	2,386

Source: Loeb, Soland, and Fox (2012)

Table 3

Stability of Teacher Value-Added Between English Learners and Others (Math)

	Ranking for Non-English Learners (Quintiles)					
Ranking for English Learners	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	50%	25%	15%	7 %	4%	467
Q2	23%	32%	23%	14%	5%	480
Q3	15%	23%	28%	21%	12%	484
Q4	9%	15%	25%	30%	20%	480
Q5 (Top)	4%	4 %	10%	28%	59%	475
Total	409	525	541	504	407	2,386

Source: Loeb, Soland, and Fox (2012)

Table 3

Stability of Teacher Value-Added Between English Learners and Others (Math)

	Ranking for Non-English Learners (Quintiles)					
Ranking for English Learners	Q1 (Bottom)	Q2	Q3	Q4	Q5 (Top)	*Total*
Q1 (Bottom)	50%	25%	15%	7 %	4%	467
Q2	23%	32%	23%	14%	5%	480
Q3	15%	23%	28%	21%	12%	484
Q4	9%	15%	25%	30%	20%	480
Q5 (Top)	4%	4 %	10%	28%	59%	475
Total	409	525	541	504	407	2,386

Source: Loeb, Soland, and Fox (2012)

What More Needs To Be Known On This Issue?

- Only a small number of studies on stability
- Future Research
  - Better separate measurement error from true differences
    - e.g., error can come from many sources (tests, small student sample, summer learning, changing school supports)
  - Systematically compare estimates across model specifications and data sets
  - Identify and assess dimensions of time, topic, and student populations
    - e.g., each subject has many topics…
  - Provide evidence on the sources of instability
    - e.g., grade changes, school changes, student changes…
    - Understanding causes of inconsistency sheds light on appropriate use of measures, and on teacher improvement

What Can't Be Resolved By Empirical Evidence On This Issue?

- Value-added measures will never be completely stable
- Choices given instability
  - Number of years of data to combine
  - Subject areas or topics to measure
  - Student groups among which to distinguish
- More generally, the optimal use of measures

Practical Implications

- Instability across years calls for
  - Caution when making decisions for which there are no mechanisms for re-evaluation and no other sources of information
  - Benefit of using multiple years of data and multiple sources of information
- Instability across subjects calls for
  - Testing valued skills and knowledge
  - Purposeful assignment of teachers to courses
- Instability/stability across student groups calls for
  - Benefit of assuring generally effective teachers for all student groups
  - Targeted professional development for differences that are evident
- Instability, more generally, suggests
  - Benefit of updating decisions with new information.

CENTER FOR EDUCATION POLICY ANALYSIS at STANFORD UNIVERSITY

cepa

How Stable Are Value-added Estimates Across Years, Subjects And Student Groups?

Susanna Loeb

Christopher A. Candelaria

Stanford University