The Use of Value-Added in Teacher Evaluations

AFT TEACH Conference
July 2015
Washington, D.C.

Matthew Di Carlo, Ph.D.
Senior Fellow
Albert Shanker Institute

ALBERT SHANKER INSTITUTE

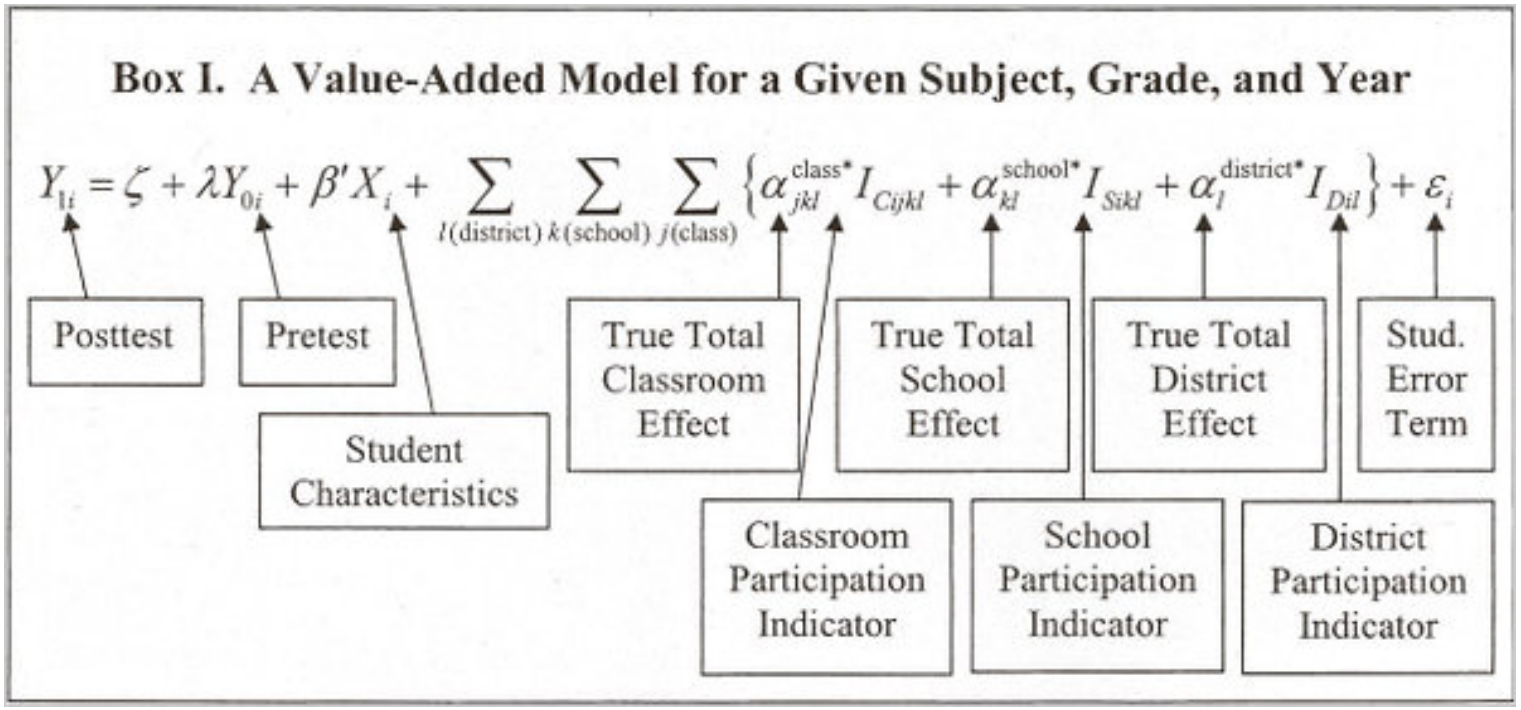
Framing points

- VA gets most of the attention in debate, but in reality a minority component for a minority of teachers (for now, at least)
- VA has many useful policy and research applications; must be separated from debate over accountability use
- Very little evidence on how to use VA in evaluations or impact of doing so
- There are different types of growth models – generalize with caution

Basic features

- Focus on progress of students, not level (unlike NCLB)
- Set expectations for student growth using observable characteristics, most important of which is prior performance
- Teachers' VA based on whether their students exceed those expectations

The scary model



The *NY Times* published this equation in 2011, and it became a symbol of value-added's inaccessibility and reductionism

VA is complex, but so is teaching and learning

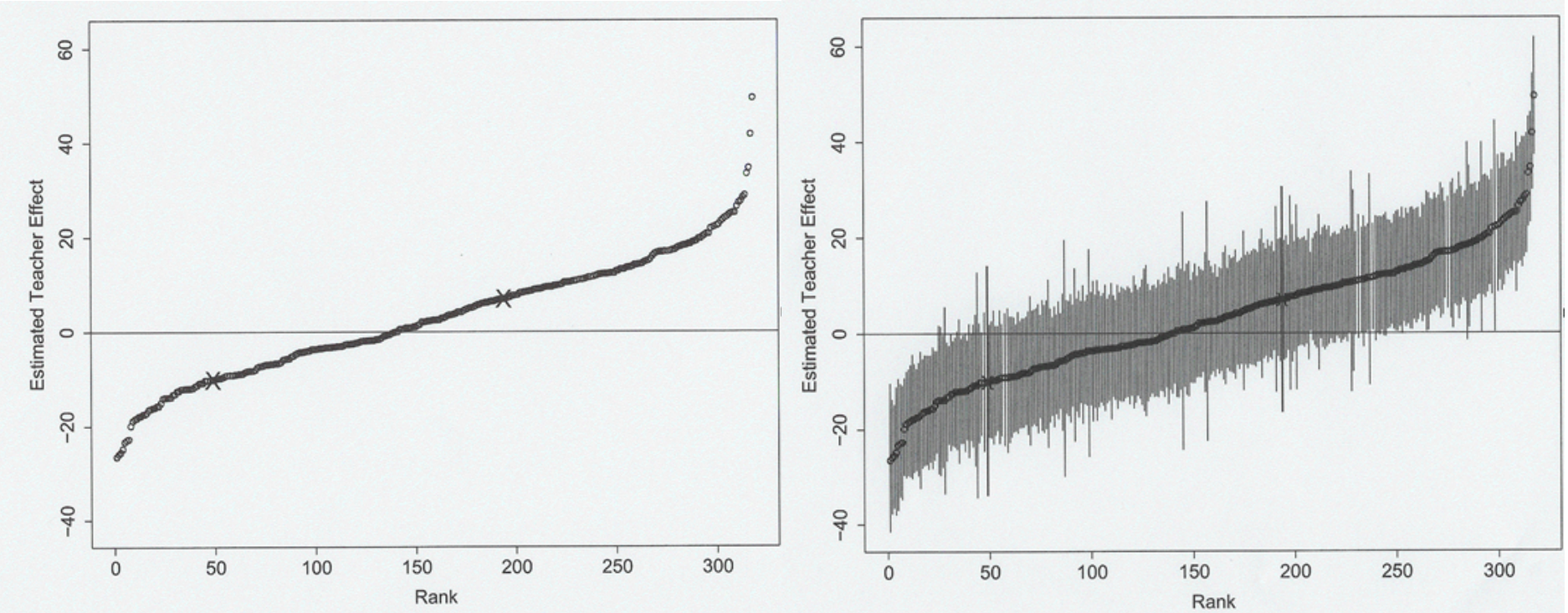
Three premises

1. Teachers should be held accountable for their job performance
2. No measure is perfect – there *will* be mistakes
3. Any measure must be assessed relative to available alternatives

Criticism 1: Unreliable

- Due largely to test measurement error and especially small samples (classes), VA estimates are "noisy"
- That is, teachers' scores are estimated imprecisely, and thus fluctuate between years
- This random error plagues virtually all school accountability systems
- May generate classification errors, as well as consequences for teacher recruitment, retention and other behaviors

Error within years



- VA scores for individual teachers, sorted
- "Average teacher" line in middle
- Error bars (right) show most teachers are "statistically average," but "truth" more likely in middle than at the ends

Adapted from: McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., and Hamilton, L.S. 2004. Evaluating Value-Added Models for Teacher Accountability. Santa Monica, CA: RAND Corporation.

Stability between years

YEAR ONE QUINTILE	YEAR TWO QUINTILE 1	2	3	4	5
1	4.2%	5.2%	5.2%	2.3%	2.9%
2	3.3%	4.2%	5.2%	4.9%	2.0%
3	2.3%	3.6%	5.2%	5.9%	3.3%
4	1.3%	2.6%	4.2%	6.5%	4.6%
5	2.3%	2.0%	2.9%	6.9%	6.9%

Source: McCaffrey, D.S., Sass, T.R., Lockwood, J.R., and Mihaly, K. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4), 572-606.

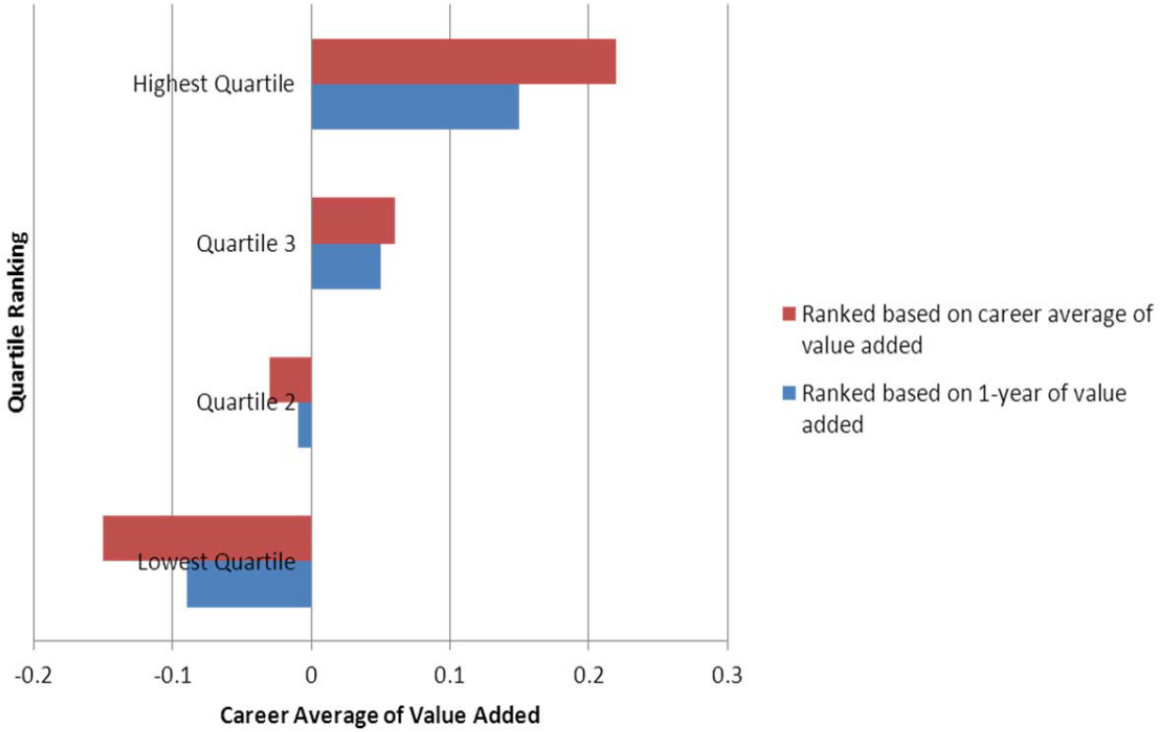
Stable	27.0%
Move 1	38.9%
Move 2	21.2%
Move 3-4	12.8%

34% of teachers moved at least two quintiles between years, while 27% remained "stable"

Clarifying reliability

- Even a perfectly unbiased measure would produce imprecise estimates, and a perfectly reliable measure is not necessarily a good one (indeed, probably is not)
- Some of the instability between years is “real” change – performance is not fixed
- Classroom observations also exhibit instability between years (in part for the same reason)

Signal : Noise



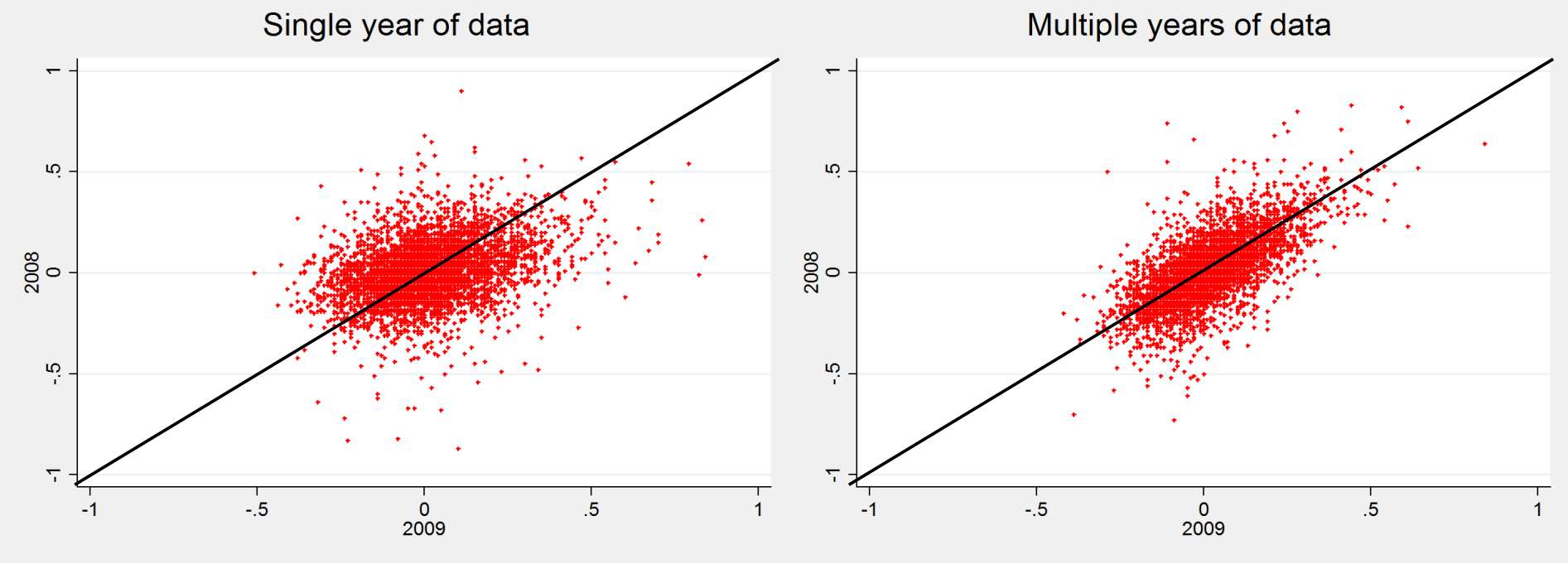
Source: Staiger, D.O. and Kane, T.J. 2014. Making Decisions with Imprecise Performance Measures: The Relationship Between Annual Student Achievement Gains and a Teacher's Career Value-Added. In Thomas J. Kane, Kerri A. Kerr and Robert C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (p. 144-169). San Francisco, CA: Jossey-Bass.

- These correlations are modest, but not random
- Simple year-to-year relationships usually range from 0.2-0.5
- And, from a longer term perspective, year-to-career correlations amy be in the 0.5-0.8 range
- Remember also that random error limits strength of year-to-year correlation even if model is perfect

The War on Error

- Random error is inevitable and a big problem for high stakes accountability use of teacher VA
- The imprecision, however, is not a feature of VA *per se*, and can be *partially* mitigated via policy design
- Addressing error entails trade offs, but may offer benefits in terms of both "accuracy" and, perhaps, perceived fairness

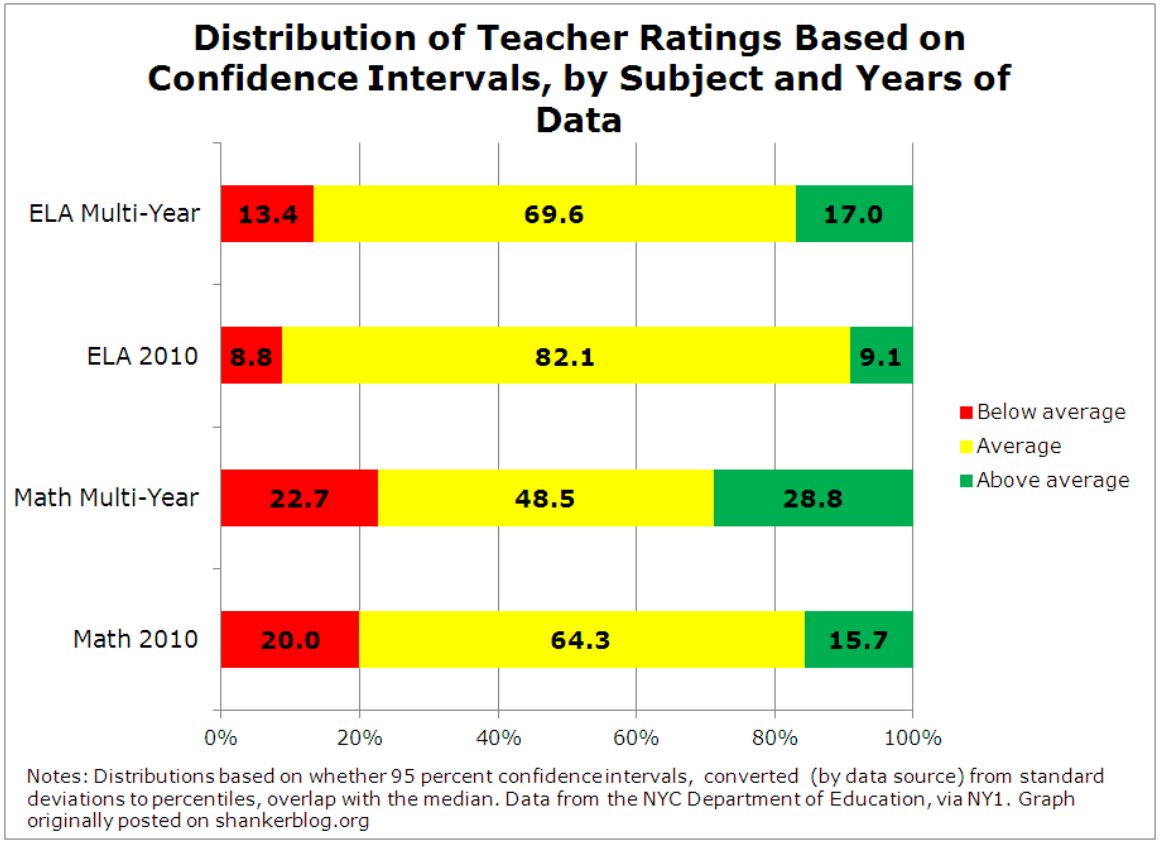
Increase sample size

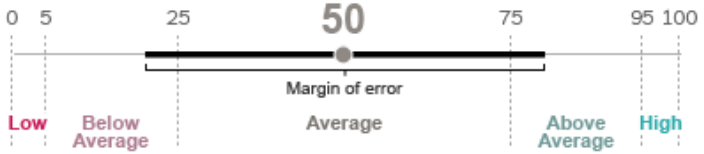


Source: Author's calculations using data from NYC Teacher Data Records

- Using multiple years of data substantially improves the stability between years – this can be done as a requirement (at least 2 years of data) or as option (2 years when possible)
- Downsides here include loss of ability to detect year-to-year variation, and possible restricting of "eligible" sample (if multiple years required)
- Statistical technique called "shrinking" estimates is a related option

Consider error margins





- It varies by subject and years of data, but most teachers' estimates are "statistically average"
- In policy context, this statistical interpretation potentially useful information – e.g., when "converting" VA estimates to evaluation scores
- Downsides here include forfeiture of information and simplicity/ accessibility

Criticism 2: Invalid

- In the "technical" sense, validity of VA is about whether models provide unbiased causal estimates of test-based effectiveness
- Students are not randomly assigned to classes and schools, and estimates biased by unobserved differences between students in different classes, as well as, perhaps, peer effects, school resources, etc.
  - Particularly challenging in high schools (e.g., tracking), and among special education teachers
- In addition, using a more expansive notion of validity, VA estimates:
  - Vary by subject, grade, and test
  - Only modestly correlated with other measures, such as observations

Variation by students

Average Math Percentile Ranks for Typical Classrooms			
Model type	**Advantaged**	**Average**	**Disadvantaged**
MGP	60.2	49.9	42.1
Lagged score VAM	64.5	50.6	39.3
Student Background VAM	57.7	50.2	47.7
Student FE VAM	51.6	47.8	48.8

Source: Goldhaber, D., Walch, J., and Gabele, B. 2014. Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. *Statistics and Public Policy* 1(1), 28-39.

- Average teacher VA percentile rank substantially lower in classrooms comprised of disadvantaged versus advantaged students
- Notice, though, that relationship varies substantially by model

Inter-measure "match"

Table 4

MET Project Correlations Between Value-Added Model (VAM) Scores and Classroom Observations

Subject area	Classroom observation system	Correlation of overall quality rating with prior year VAM score
Mathematics	CLASS	0.18
Mathematics	FFT	0.13
Mathematics	UTOP	0.27
Mathematics	MQI	0.09
English language arts	CLASS	0.08
English language arts	FFT	0.07
English language arts	PLATO	0.06

Note: Data are from the MET Project (2012, pp. 46, 53). CLASS = Classroom Assessment Scoring System, FFT = Framework for Teaching, PLATO = Protocol for Language Arts Teaching Observations, MQI = Mathematical Quality of Instruction, UTOP = UTeach Teacher Observation Protocol.

Source: MET project summarized in: Haertel, E.H. 2013. *Reliability and Validity of Inferences About Teachers Based on Student Test Scores*. Princeton, NJ: Educational Testing Service.

- This is a broader notion of validity, but value-added scores are a rather weak predictor of observation scores, particularly in ELA, and regardless of protocol
- This *may* suggest that VA is not strongly related to instructional quality, and that estimates vary for reasons other than what teachers actually do in the classroom

Clarifying validity

- Validity is a feature of how measures are interpreted, not measures themselves
- There is some disagreement about extent of bias in VA estimates, and within- versus between schools an important distinction (but there will be individual teachers affected regardless of extent)
- Association between VA and long term student outcomes[1]
- There is no reason to expect (or perhaps even want) VA to match up with other measures
- Association between VA and student/school characteristics varies substantially by model, and some of it is "real"
- Also →

1 Chetty, R., Freidman, J.N., and Rockoff, J.E. 2014. Measuring the Impacts of Teachers I & II. *American Economic Review* 104(9), 2593-79.

Other measures & FRL

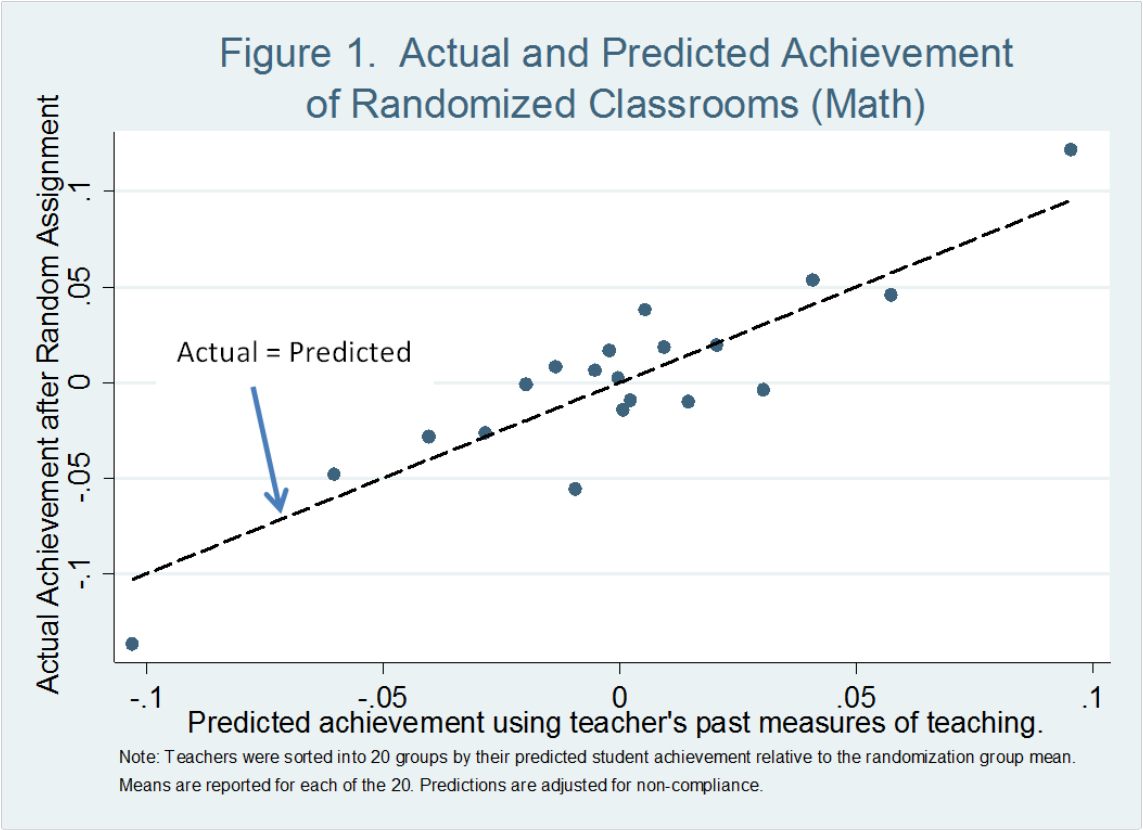
Minneapolis Public Schools Teacher Evaluation Results, by Component, 2013-14

Ranking	Math value-added		Reading value-added		Classroom observations		Student surveys	
	Avg. FRL rate	# Schools	Avg. FRL rate	# Schools	Avg. FRL rate	# Schools	Avg. FRL rate	# Schools
Far above average							62.3	2
Above average	57.1	5	51.7	6	42.9	15	64.8	13
Average	64.1	25	64.4	22	67.2	24	69.6	26
Below average	74.8	10	82.3	9	86.1	14	73.8	9
Far below average							44.1	3

Notes: FRL averages weighted by teacher frequency. Data source: Minneapolis Star Tribune

- In most other places with new evaluations, clear relationship between student characteristics and *all* non-VA measures, including classroom observations and student surveys
- To the degree association represents "bias," all measures being used in new systems exhibit it

Experimental evidence



Source: Kane, T.J., and Staiger, D.O. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper 14607.

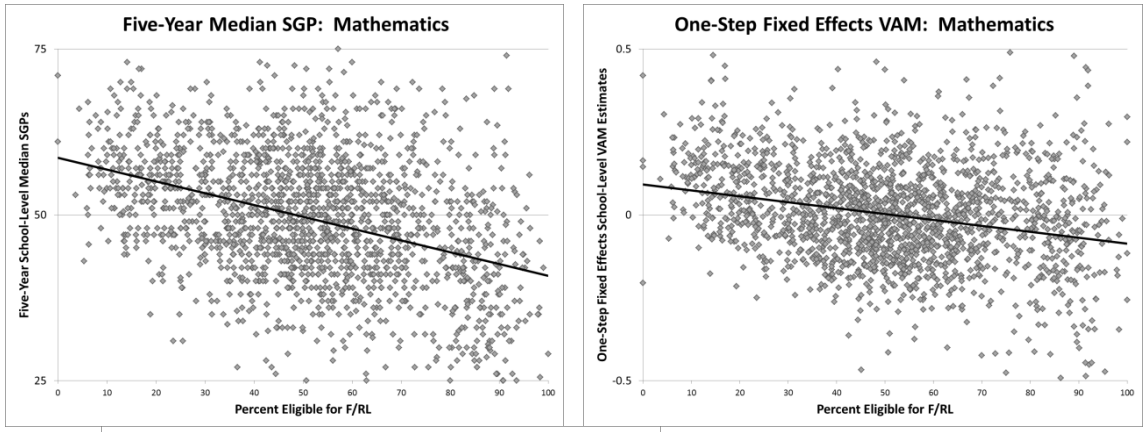
- A few studies have assessed VA under random classroom assignment
- In one such study, estimates under random sorting in year one generally consistent with random sorting estimates in year two
- Does not, however, preclude individual errors, and may not hold up for all teachers in all districts

Validation nation

- Cannot observe “true” teacher performance - we know there are at least some mistakes, but not necessarily how to identify them
- VA need not be perfectly unbiased causal estimate to be useful, and perfectly unbiased estimates would still generate misclassifications (e.g., random error)
- Nevertheless, concerns about validity very important, and impossible to eliminate; states and districts should be doing more to assess and monitor (including, by the way, roster verification)
- Also, perhaps, some leverage in policy design

Model choice

Figure 2. School Growth Measures from Each Model Plotted Against School Shares Eligible for Free/Reduced-Price Lunch.



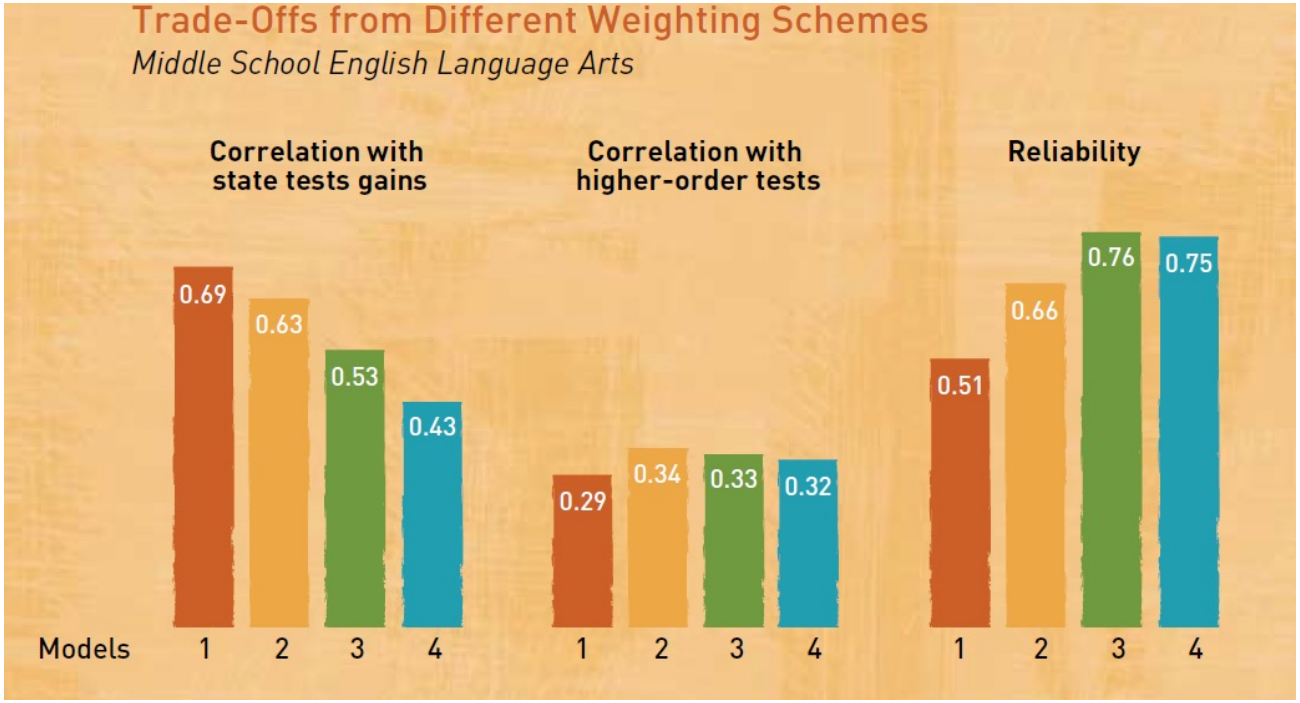
Correlation: -0.37 Correlation: -0.25 Correlation: - 0.03

Using two-step FE model, correlation is zero

- Relationship between VA and student characteristics can vary substantially by the type of model used
- These are school estimates, but the same goes for teacher VA
- Note that USED guidelines discourage using some control variables

Source: Ehlert, M., Koedel, C., Parson, E., and Podgursky, M. Forthcoming. Selecting Growth Models for School and Teacher Evaluations: Should Proportionality Matter? *Education Finance and Policy*.

Weighting judgment



Source: The Measures of Effective Teaching Project

In this figure, model 1 weights value-added most heavily, and model 4 least heavily, vis-à-vis classroom observations and student surveys

- Everyone supports "multiple measures," but the choice of weight for VA (or any measure) primarily a *value judgment*
- Weighting any outcome more will predict that outcome better and the others worse
- Different weights can also lead to different incentives

Weighting alternatives

- The simple weighting systems used by most states are not the only option (at least in theory)
- Some researchers have proposed using VA as a sort of "screening device," by which teachers are identified for further observation and remediation[1]
- This type of alternative approach might better exploit strengths and weaknesses of VA (and those of other measures), thus improving both reliability and validity of inferences

1 For example, see: Harris, D.N. 2012. Creating a Valid Process for Using Teacher Value-Added Measures (Shanker Blog post 11/28/12). Washington, D.C: Albert Shanker Institute.

Criticism 3: Bad Incentives

- In the final analysis, the important outcome is whether using VA in evaluations improves outcomes, and that depends largely on how current and prospective teachers respond
- Even if perfect, VA may not have the desired effect, and may cause harm
  - Does not help teachers improve
  - Encourages “teaching (or “principaling”) to the test
  - Disincentive to teach in high needs schools/classrooms
  - Adverse impact on labor supply
- These are all empirical questions, but there is relatively little evidence

The teacher factor

- We don't know how teachers will respond, and it will vary within and between locations
- Some common sense suggestions:
  - Monitor attitudes/behavior every year
  - Avoid schoolwide VA (apparently)
  - Take time for piloting and implementation
  - Build research design into policy
  - Calibrate risks and rewards
- Even the most well designed evaluations will not work if they don't change behavior, and they may have negative impact

Review

- Address random error with sample size requirements, shrinkage, and/or “conversion”
- Consider models that mitigate association between estimates and student characteristics
- Weight your judgment (or try a different approach)
- Monitor attitudes and behavior every year
- Take your time
- Build in research assessment
- Calibrate risks and rewards

Additional reading

- There are countless resources available. A few of note:
  - Harris, D. 2011. *Value-Added Measures in Education: What Every Educator Needs to Know*. Cambridge, MA: Harvard Education Press.
  - Carnegie Knowledge Network (various authors). *Value-Added Knowledge Briefs* (15 briefs). Washington, DC: Carnegie Knowledge Network.
  - Haertel, E. H. 2013. *Reliability and Validity of Inferences About Teachers Based on Student Test Scores*. Princeton, NJ: Educational Testing Service.
  - Koedel, C., Mihaly, K., and Rockoff, J. Forthcoming. Value-Added Modeling: A Review. *Education Finance and Policy*. (More technical.)

Closing arguments

- The “technical” problems regarding VA are very important, but can be at least partially addressed via policy design (latter is not happening in many places)
- The true purpose of accountability systems, however, is to change or reinforce behavior, and so effect of VA in evaluations will depend on how teachers respond
- This impact will be difficult to anticipate and measure
- The only unsupportable position at this point is certainty

Thank You

Matthew Di Carlo
mdicarlo@ashankerinst.org
Albert Shanker Institute

ALBERT SHANKER INSTITUTE