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
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 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
Criticisms 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
The Use of Value-Added in Teacher Evaluations

- 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

Stability between years

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

• 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 may 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
The Use of Value-Added in Teacher Evaluations

- 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

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

<table>
<thead>
<tr>
<th>Model type</th>
<th>Advantaged</th>
<th>Average</th>
<th>Disadvantaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGP</td>
<td>60.2</td>
<td>49.9</td>
<td>42.1</td>
</tr>
<tr>
<td>Lagged score VAM</td>
<td>64.5</td>
<td>50.6</td>
<td>39.3</td>
</tr>
<tr>
<td>Student Background VAM</td>
<td>57.7</td>
<td>50.2</td>
<td>47.7</td>
</tr>
<tr>
<td>Student FE VAM</td>
<td>51.6</td>
<td>47.8</td>
<td>48.8</td>
</tr>
</tbody>
</table>


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

Inter-measure “match”

- 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.

Table 4
MET Project Correlations Between Value-Added Model (VAM) Scores and Classroom Observations

<table>
<thead>
<tr>
<th>Subject area</th>
<th>Classroom observation system</th>
<th>Correlation of overall quality rating with prior year VAM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>CLASS</td>
<td>0.18</td>
</tr>
<tr>
<td>Mathematics</td>
<td>FFT</td>
<td>0.13</td>
</tr>
<tr>
<td>Mathematics</td>
<td>UTOP</td>
<td>0.27</td>
</tr>
<tr>
<td>Mathematics</td>
<td>MQI</td>
<td>0.09</td>
</tr>
<tr>
<td>English language arts</td>
<td>CLASS</td>
<td>0.08</td>
</tr>
<tr>
<td>English language arts</td>
<td>FFT</td>
<td>0.07</td>
</tr>
<tr>
<td>English language arts</td>
<td>PLATO</td>
<td>0.06</td>
</tr>
</tbody>
</table>

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.

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”.

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

- 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.

- 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

Using two-step FE model, correlation is zero

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.

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

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:
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