How Non-Linearity and Grade-level Differences Complicate the Validation of Observation Protocols

Valeriy Lazarev and Denis Newman
Empirical Education Inc.
Outline

1. Motivation: teacher evaluation systems
2. Issues: non-linearity and variability
3. Approach: generalized additive models
4. Results
5. Discussion: implications
Motivation

• Teacher evaluation is currently a major policy issue, driven in large part by DoE requirements. However, measures of teacher effectiveness used for purposes of evaluation are still immature.

• Components of evaluation systems (including test-based metrics of student growth and observation-based metrics of teacher performance) need to be validated (analyzed for consistency and relevance) before they can be used for high-stakes decisions.

• In particular, we need to understand how metrics arising from systematic classroom observations related to other metrics of student growth.
Related work

• Technical assistance to school systems that are concerned with the validity of the observational measures incorporated into teacher evaluation

• Experimental research where the program impact on student achievement is expected to be mediated by a measure of classroom practice
Related work
Measures of Effective Teaching (MET)

• Massive data collection and analysis of relationship between aggregate measures of classroom practice and student achievement.
• In the view of MET studies, instruments (observation rubrics and student surveys) can reveal the latent teacher effectiveness and are therefore correlated with a given student growth metric, which is the ultimate “output” of teaching practice.
Related work
Measures of Effective Teaching (MET)

• Evidence of significant but weak positive correlation between various metrics of teacher performance (Kane and Staiger, 2012). The problem may lie with the robustness of value-added estimates.

• Little attention to functional relationships:
  – Observation scores are not necessarily linear indicators of TVA.
  – Expectation that observation instruments perform the same way in every classroom may be unrealistic given the variability of teaching practice across subjects and grade levels.

⇒ Low reported (linear) correlations, which are inadequate metrics.
Study objective

- We explore the patterns of relationship between observational scores and value-added measures of teacher performance and the variation in these relationships across grades 4 - 8.
- We focus on the developmental changes in what can be described as effective student-teacher interaction as children progress from elementary to middle school.
- We treat each component separately (as opposed to using averages of component scores as in earlier studies) since each one is intended to measure a separate aspect of classroom practice.
Research questions

• Do the relationships between observation scores and their value-added scores tend to be non-linear?
• Are these relationships affected by developmental changes as indicated by grade levels?
• To what extent are the observed changes associated with increasing departmentalization (in upper elementary and middle school)?
• Are these differences seen in both math and English language arts?
Approach

- Observation component scores are different and imperfect instruments that can be used to measure teacher effectiveness, with appropriate calibration.
- Readings of a good instrument do not have to be a linear function of the measured characteristic but it has to be monotonic.
- We do not make an assumption that observational protocols are “generic” (Danielson, 2007) - apply equally to teaching any content at any grade level.
Approach

• Technically: Observation component score is an arbitrary function of the value-added score. Our primary goal is to establish the “true” shape of such a function.

⇒ Nonparametric approach: estimation of a generalized additive model using penalized regression splines (Wood 2006); allows identifying the true shape of the relationship.

• Basic model used for exploratory analysis: \( y_i = s(x_i) + \varepsilon_i \), where \( x \) - student growth metric, \( y \) - observational score, \( s(x) = E(y \mid x) \) – an unknown smooth function, and \( \varepsilon_i \) is the zero-mean random error term.
Approach:
Departmentalization vs. gradual developmental change

• If departmentalization is the main factor responsible for the different shapes of functional relationships across grades, then
  \[ y_i = s_{nd}(x_i) + s_d(x_i) + \varepsilon_i \] - separate functions are estimated for non-departmentalized and departmentalized teaching setting.

• If the observed change is due to developmental factors, then the relationship is changing gradually with grade level. An appropriate model in this case is a two-dimensional smooth function allowing for arbitrary interactions with the grade-level:
  \[ y_i = s_2(x_i, g_i) + \varepsilon_i \]
  where \( g_i \) is grade level.

• To analyze the relative contribution of the two types of factors, we estimate a single nested model:
  \[ y_i = s_{nd}(x_i) + s_d(x_i) + s_2(x_i, g_i) + \varepsilon_i \]
  This model is identifiable as long as data on departmentalized elementary school classrooms are available.
Data

- Upper-elementary and middle-school math and ELA teachers assessed on multiple metrics by MET project.
- Observation: Framework for Teaching rubric (eight components of two domains: “Classroom environment” and “Instruction”).
- Value-added scores: based on study-administered BAM (math) and SAT9 (ELA) tests.
Data

- Sample:
  - 1102 usable data points
  - Grade levels: 4-8.
  - Departmentalized teaching: 25% of elementary school teachers, 95% of 6th grade teachers, and all 7th and 8th grade teachers.
- Diverse student population (class characteristics not included in the analysis)
Results

Non-linear (and non-monotonic) relationships between observation scores and value-added scores

Observation scores are normalized.
Results

Relationships between observation scores and value-added scores vary by subject

Component “Communicating with students.” Left pane: Math. Right pane: ELA. Observation scores are normalized.
Results

Relationships between observation scores and value-added scores vary by grade level

## Results

Statistical significance of interaction terms. Math

<table>
<thead>
<tr>
<th>FFT Component</th>
<th>R²</th>
<th>Departmentalized teaching, p value</th>
<th>Grade level, p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating an Environment of Respect and Rapport</td>
<td>0.093</td>
<td>0.14</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Communicating With Students</td>
<td>0.076</td>
<td>0.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Establishing a Culture for Learning</td>
<td>0.134</td>
<td>0.05</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Engaging Students in Learning</td>
<td>0.095</td>
<td>0.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Managing Classroom Procedures</td>
<td>0.071</td>
<td>0.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Managing Student Behavior</td>
<td>0.077</td>
<td>0.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Using Assessment in Instruction</td>
<td>0.069</td>
<td>0.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Using Questioning and Discussion Techniques</td>
<td>0.048</td>
<td>0.57</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
## Results

Statistical significance of interaction terms. ELA

<table>
<thead>
<tr>
<th>FFT Component</th>
<th>R²</th>
<th>Departmentalized teaching, p value</th>
<th>Grade level, p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating an Environment of Respect and Rapport</td>
<td>0.095</td>
<td>0.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Communicating With Students</td>
<td>0.074</td>
<td>0.08</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Establishing a Culture for Learning</td>
<td>0.142</td>
<td>0.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Engaging Students in Learning</td>
<td>0.099</td>
<td>0.21</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Managing Classroom Procedures</td>
<td>0.108</td>
<td>0.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Managing Student Behavior</td>
<td>0.087</td>
<td>0.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Using Assessment in Instruction</td>
<td>0.091</td>
<td>0.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Using Questioning and Discussion Techniques</td>
<td>0.088</td>
<td>0.16</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
Results

Fine structure of relationships between observation scores and value-added. Ex. 1.

Results
Fine structure of relationships between observation scores and value-added. Ex. 2.

Conclusion

- Observation-student growth relationships:
  - Are non-linear
  - Differ between math and ELA
  - Vary with grade level - likely a gradual change associated with developmental rather than organizational factors (such as departmentalized teaching)
Implications

• For effectiveness research and “instrumentation”:
  – analysis should not be limited to correlations ⇒ more data and more complex models are needed
  – components cannot be unambiguously associated with achievement gains ⇒ are they measuring other latent factors?

• For teacher evaluation systems:
  – composite scoring formulas should account for non-linearities, simple sums or averages may not be adequate
  – classroom observation instruments should be calibrated for use in particular types of teaching environments
Thank you

Val Lazarev, Denis Newman

Empirical Education Inc.

vwww.empiricaleducation.com