Challenges with the sensitivity of value-added assessment measures of teacher performance

Leslie R. Hawley
James A. Bovaird
Chaorong Wu

Nebraska Center for Research on Children, Youth, Families and Schools
Introduction

- So far we have focused our discussion on the ability of a test (based on test-takers’ scores) to accurately reflect the quality of instruction.

- However, there are additional issues to consider with the use of test scores to evaluate teachers (and/or principals, schools).

- Even if evidence were provided for the ability of test scores to accurately reflect instructor performance, the methods used to evaluate teachers are still in question:
  - Namely, there are questions as to the validity of inferences from value-added assessment methods.
Value-added assessment (VAA)

- Statistical analysis in which longitudinal data are used to study the effects of inputs (teachers) on an outcome measure (educational achievement) (McCaffrey & Lockwood, 2008)
  - Uses (or proposed uses) include but are not limited to professional development, promotion, tenure, and retention

- Goal is to estimate the contribution teachers make to learning
  - Did students grow more or less than expected?
Value-added assessment (VAA)

- Predicted achievement is based on a typical school with similar students
- In the example below, actual achievement is higher than predicted so a teacher’s value-added is higher than average

(Adapted from Harris, 2011)
Value-added assessment (VAA)

- Value-added piece of any of the statistical models is dictated by the use or proposed use rather than any explicit model specification (Betebenner, Wenning, & Briggs, 2011)

- Several types of VAA models
  - Models range from simple to complex
  - Usually incorporate additional variables to adjust for factors outside the control of the school/teacher
    - Education Value-Added Assessment System (EVAAS) relies primarily on test scores
  - Models are not standardized, so model choice depends on the combination of contractor (researcher/company) and client (state/district)
A sample of potential challenges with the sensitivity of VAA models

- Random assignment
- Inclusion (or exclusion) of covariates
- Missing data
Random Assignment

- Potential for bias in VAA estimates due to non-random assignment of students to teachers (schools)
  - Any non-random assignment (sorting/tracking) is theoretically controlled for by the inclusion of covariates (e.g., prior scores, SES measures) or in the case of some models (i.e., EVAAS), students act as their own controls
- Sorting effects
  - Rothstein’s falsification test (2010)
  - Similar results found by Briggs and Domingue (2011)

- A handful of experimental research has been conducted but these studies are not a panacea
  - Kane & Staiger (2008)
  - Measures of Effective Teaching (MET) project (2010 - 2013)
Covariates

- Additional variables are can be included to control for non-school factors which may influence test scores
  - “Ensures that each teacher is in effect compared only against other teachers with demographically similar classrooms” (Gordon et al., 2006, p. 20)
  - Variables can include but are not limited to a variety of student-, peer-, and/or school-level variables

- Inclusion of covariates and the types of variables is more than a statistical decision
  - May want to control for possible non-school related factors influencing achievement but do not want to create (or give the appearance of) different standards
Covariates

Research has yielded mixed results on the inclusion (or exclusion) of covariates:

- Models are more sensitive to the choice of outcome measure than specifications (e.g., covariates) (Lockwood et al., 2007; Papay, 2011)

- Model estimates and/or teacher rankings are sensitive to the inclusion (or exclusion) of covariates (Briggs & Domingue, 2011; McCaffrey et al., 2004; Tekwe et al., 2004)

- Omitting student characteristics in EVAAS model that accounts for several years of growth does not appear to bias teacher estimates (Ballou, 2005)
Missing data

• Models tend to be sensitive to the amount of data available
  • Wider confidence intervals with 1 year compared to 3 years of information (Corcoran, 2010)
  • Error rates for teacher-level analyses will be about 26% with 3 years of data (1 in 4 are incorrectly classified) (Schocet & Chiang, 2013)

• Missing data are the norm rather than the exception in student data systems
  • Not uncommon to find 30 – 40 percent of relevant student data are missing (Braun & Wainer, 2007; Lockwood, 2006)
Missing data

- Availability of longitudinal data influences what information can be included in the models and which teachers receive VAA estimates
  - Gain scores require two consecutive years of data
  - Models that include a large number of covariates

- Some more sophisticated modeling approaches can handle missing data but statistical methods may not always be able to compensate
  - For instance, the combination of small class size and missing data would likely impact the degree of error and stability in VAA estimates

- Additional concerns with the statistical assumptions regarding the pattern of missing data
  - Assumptions may not hold if certain students are more likely to be missing information
Covariates & Missing data—Example

• Small simulation study (pilot data for project)

• Population:
  • Three time points; linear growth; no teacher effect persistence
  • 50 teachers per grade; 25 students per teacher
  • 2 covariates related to achievement (different degrees of relationship)
  • No missing data & 50% randomly missing

• Analyzed population sample with a cross-classified missed effects model across six different scenarios
  1) No missing data; No covariates
  2) No missing data; 1 covariate
  3) No missing data; 2 covariates
  4) Missing data; No covariates
  5) Missing data; 1 covariate
  6) Missing data; 2 covariates
Covariates & Missing data—Example

Teacher VAA estimates were converted into rankings.

Spearman rank order correlations were calculated between the true teacher rank and the estimated teacher rank.

<table>
<thead>
<tr>
<th>Rank Order Correlation between True Teacher Rank and Simulation Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Missing Data</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Time 1</td>
</tr>
<tr>
<td>Time 2</td>
</tr>
<tr>
<td>Time 3</td>
</tr>
</tbody>
</table>
Covariates & Missing data—Example

Time 3 Teacher Ranks: No missing data

Time 3 Teacher Ranks: Missing 50% of data
Discussion

- A defensible accountability system based on students’ achievement scores requires valid and reliable measures as well as well-defined, systematic procedures for analyzing achievement data (Linn, 2005)

- Presented information is a limited example of the sensitivity of VAA teacher estimates
  - Other potential aspects that may influence the stability of estimates/ranks include but are not limited to:
    - Small class sizes
    - Special populations, English language learners
    - Red Sox Phenomenon (2011 season)

- These factors on their own are challenges with VAA systems and they interact in complex ecosystems
Potential application

- Professional development
  - Statistical measures could shed light on which teachers need additional help
  - But where do you start?

- “But even the best models are not pure measures of teacher effectiveness. VAM scores do predict important student learning outcomes, but my reading of the evidence strongly suggests that these scores are nonetheless measure not only how well teachers teach, but also whom and where they teach.” (Haertel, 2013, p. 17)
Potential application

- Multiple measures of performance
  - Multiple indicators can help mitigate potential sources of error
    - MET study (2013) has explored the combination of observations, student surveys, and value-added scores

- Additional questions with using multiple measures approach:
  - Weight given to value-added estimates over other indicators?
  - Which is the gold standard?

- “Teacher VAM scores should emphatically not be included as a substantial factor with a fixed weight in consequential teacher personnel decisions.” (Haertel, 2013, p. 23)
Thank you.
lhawley2@unl.edu