Some Lessons Learned from Our (and Others') VAM Research

Jeffrey M. Wooldridge
Department of Economics
Michigan State University

October 10, 2013

These comments report on research supported in part by a Pre-Doctoral Training Grant from the Institute for Education Sciences, U.S. Department of Education (Award #R305B090011) to Michigan State University and an IES Statistical Research and Methodology Grant (Award #R305D100028).

1. Estimation

- Many different approaches to estimating teacher value-added are being used.
  - Estimators based on educational production function models, such as the cumulative effects model.
  - Estimators derived more from an astructural "treatment effects" perspective.
  - Some estimators can be given both interpretations.

- Guarino, Reckase, and Wooldridge (forthcoming, EFP) find that no estimator works well uniformly across student grouping and student assignment mechanisms.
    - Structural-based estimators, such as the Arellano and Bond dynamic panel data estimator, do not work especially well for estimating VAMs.
    - Standard fixed effects estimators are very sensitive to dynamic grouping mechanisms, but are occasionally best when assignment is based on unobservable student characteristics.

- Regression estimators that control for past test scores can perform well even when based on a "misspecified" model (student heterogeneity, neglected serial correlation, ceiling effects).
- Even "dynamic OLS" is not a panacea, especially when assignment is based on unobserved heterogeneity.

- Several estimation approaches do not allow teacher assignment to be correlated with past test outcomes. ("Random" versus "Fixed" teacher effects.)
  - Empirical Bayes'
    - But with many students per teacher, EB approaches DOLS.
  - "Average Residual" Methods
  - Percentile Growth Methods

- Consider a standard linear model:

$$y = x_1\beta_1 + x_2\beta_2 + u$$

We are interested in β_1 (teacher effects). x_2 includes controls, such as past test scores and demographics.

- OLS can be obtained by first regressing x_1 on x_2 and getting the residuals, say r_1. Then, regress y on r_1.
  - We can also first regress y on x_2 to get residuals r_0 and then regress r_0 on r_1.

- It is not generally valid to regress r_0 on x_1 because this does not allow for correlation between x_1 and x_2.
  - In effect, AR and CGM do not allow for correlation between teacher assignment and the controls.
  - If shrinkage is desired, can apply shrinkage factors to OLS estimates that account for endogenous teacher assignment.
- For CGM, need to determine how to control for nonrandom assignment.

- In simulations, methods that don't properly partial out can work poorly under nonrandom teacher assignment. The "shrinkage" does not help with nonrandom assignment.
  - How important is the misspecification empirically?

    Not very if teacher assignment is close to random.
- Any estimator will be inconsistent in certain scenarios. But how do they work in practice? Estimators should be evaluated via careful simulation studies using plausible, flexible models to describe the evolution of test scores.

- What kind of model would justify the AR approach? School VAMs?
- What underlying model would justify the Colorado growth model? Or, at least, when will the CGM outperform standard VAM estimators?
- And what about measurement of scores? Item Response Theory, classical measurement error. Is nonrandom assignment based on observed test scores?
- Introducing peer effects adds even further complications and may cause additional bias in VAMs.

2. Precision and Uncertainty

- Several methods proposed for obtaining standard errors for VAMs. But the setting is nonstandard: Relatively few cohorts per teacher, correlation within classrooms and schools.
- Relatively little has been done to evaluate methods for obtaining standard errors and confidence intervals of estimated VAMs.
- Standard panel data clustering at the student level works under random grouping and random assignment.

- Nonrandom grouping or tracking induces correlation within school and cohort.
- Do we have enough cohorts to justify clustering at the cohort level? A current VAM project suggests 10 cohorts could be enough. But is it realistic to be able to pool 10 cohorts per teacher?

3. Specification Testing

- Specification tests, especially the Rothstein (2010) falsification test, are often applied in VAM settings.
- Do specification tests provide any useful information? Very little for the purposes of estimating VAMs.
- An important lesson from our work: Testing assumptions in a structural model is often counterproductive for estimating teacher VAMs, especially when ranking and classifying teachers are considered most important.

- If flexible VAM estimation methods are used there is little scope for testing. Specification tests are only available when restrictions are imposed on the assignment (such as only one lag matters for teacher assignment).
- The “falsification tests” from the treatment effects literature are essentially feedback tests and are available only by imposing assumptions on tracking/assignment.

- Do we trust statistical and econometric methods to extract useful information in the context of value-added measurement?
  - More work needs to be done on assessing standard errors.
  - We need to better understand the effects of measurement error, especially under nonrandom assignment.