Evaluating Specification Tests in the Context of Value-Added Estimation

Cassandra M. Guarino
Mark D. Reckase
Brian Stacy
Jeffrey M. Wooldridge

Acknowledgment: This work was 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).

Introduction

- Value-Added models are increasingly being used to assess teacher effectiveness.
  - Nominally, different estimators appear to require different assumptions.
  - Can we determine whether the right conditions exist to accurately estimate teacher value added?
- Can tests be used as tools for choosing among estimators?
  - Unfortunately, the answer appears to be no.
  - Perhaps this should have been obvious from the beginning.

Focus of the Paper

- Explore two specification tests in the context of value-added estimation, with modifications of each that have advantages for VAM.
  - The Hausman test comparing the random effects and fixed effects estimators (where the effects are at the student level). More common in general panel data applications than VAM estimation.
  - A feedback test similar to a Rothstein (2010) "falsification test".

- Kinsler (2012) and Goldhaber and Chaplin (2012) also study versions of Rothstein's test and reach similar conclusions. Harris, Sass, and Semykina (2010) apply a battery of tests to actual data.
- Our tests come directly from the panel data literature for random effects and fixed effects estimation.
- When applied to dynamic regression, our approach is similar to tests of "unconfoundedness" from the treatment effects literature.

- Both tests try to detect nonrandom assignment of students to teachers.
  - Hausman test has power for teacher assignment depending on student heterogeneity (static assignment) as well as dynamic assignment (past test scores).
  - When applied to the FE estimator – the original intent of feedback tests in the panel data literature – the test should detect dynamic assignment but not "heterogeneity assignment."
  - Focus on the feedback test for this talk.

A General VAM Formulation

- Assume test scores are generated by

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + u_{it} - \lambda u_{i,t-1}$$

$$u_{it} = \rho u_{i,t-1} + r_{it},\ t = 1,2,\dots,T$$

  - A_{it} is achievement, E_{it} is school inputs (teacher dummies), c_i is student heterogeneity, u_{it} is an underlying structural error.
  - Can be derived from a structural cumulative effects model (CEM). Already imposes geometric decay, AR(1) serial correlation of structural errors, time homogeneity.

- For consistent estimation of the decay parameter, λ, and teacher effects, β_0, standard methods assume $\lambda = \rho$ – called a "common factor restriction" (CFR).
    - If c_i is absent, can test the CFR in the context of dynamic least squares. Presence of c_i leads to rejection even if $\lambda = \rho$.
- Important point about the CFR from GRW (forthcoming): it is not needed for DOLS to provide useful VAM estimates. So, a strong rejection can occur even if we are using a good estimation procedure.

Exogeneity Assumptions

- With the CFR imposed, consider the equation

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + r_{it}$$

  where the r_{it} are serially uncorrelated.

  - "Heterogeneity exogeneity" is when the inputs, E_{it}, are uncorrelated with c_i for all *t*.
  - "Strict exogeneity" is when E_{is} is uncorrelated with r_{it} for all *t* and *s*. Rules out feedback to future teacher assignment.

- Standard methods require strict exogeneity of inputs when the dynamic model is viewed as a structural CEM.

- Pooled OLS and random effects methods require heterogeneity exogeneity.
- Therefore, it seems sensible to test these two assumptions.
- However, if the goal is to estimate VAMs for ranking teachers, the tests may be "too powerful."
  - Strict exogeneity may not hold, but dynamic least squares can do very well.
  - Nonrandom grouping of students can cause rejection even if assignment to teachers is random.

Estimating Equations and Estimators

- The dynamic OLS (DOLS) estimator is based on

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + r_{it}$$

and so it ignores any student-specific heterogeneity.

  - Strict exogeneity of E_{it} is not needed!

- The pooled OLS (POLS) estimator sets $\lambda = 1$:

$$\Delta A_{it} = \tau + E_{it}\beta_0 + r_{it},$$

where ΔA_{it} is the gain score.

- Random Effects and Fixed Effects:

$$\Delta A_{it} = \tau + E_{it}\beta_0 + c_i + r_{it}$$

  - Both assume strict exogeneity of inputs with respect to r_{it}.
  - RE assumes heterogeneity exogeneity, so E_{it} is uncorrelated with c_i for all *t*. (A feasible GLS estimator.)
  - FE removes c_i but is inconsistent with feedback. Teacher assignment cannot react to past gain score.

Feedback or Falsification Test (Dynamic Assignment)

- Because FE actually requires strict exogeneity for consistent estimation, natural to apply test to FE.
- A regression-based test is simplest. Include the lead teacher assignments in a gain-score equation:

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + E_{i,t+1}\delta + c_i + r_{it}, t = 1, \ldots, T-1$$

  - Lose the last grade. Need three years of gain scores to implement test. If $T = 2$, nothing to test.
  - Estimate the equation by FE and compute a fully robust *F* statistic of $H_0: \delta = 0$. There are as many *df* as teachers.

- Our new proposal: Use a one *df* test that replaces all teacher dummies with the estimated VAM for next year's teacher.

- Why does the test reject, say, for FE under DG/RA? Would hope that it does not reject.
  - Within cohort/school correlation causing "cluster sampling" problems.
  - Original form of the test has too many degrees-of-freedom.
  - Knowing something about next year's teacher VAM provides information about the student's score this year.

- A higher VAM means, on average, that teacher got better students. If students are grouped by past test scores, the student is, on average, better. So the future teacher's VAM has predictive power for the current test score.

- Following Rothstein (2010), we apply the feedback test to POLS and DOLS (and RE), too, even though DOLS generally accounts for feedback.
- DOLS formulation:

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + E_{i,t+1}\delta + r_{it}$$

The Simulation Design and Results

- Closely follows Guarino, Reckase, and Wooldridge (forthcoming, EFP). Focus on nonrandom assignment within school.
- Data represent three elementary grades per student (grades 3 through 5 and an initial second-grade test score):

$$A_{i3} = \lambda A_{i2} + \beta_{i3} + c_i + e_{i3}$$

$$A_{i4} = \lambda A_{i3} + \beta_{i4} + c_i + e_{i4}$$

$$A_{i5} = \lambda A_{i4} + \beta_{i5} + c_i + e_{i5}$$

- 10 schools
- 3 grades (3^{rd}, 4^{th}, and 5^{th}) of scores and teacher assignments, with a base score in 2^{nd} grade
- 4 teachers per grade (thus 120 teachers overall)
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools
- We vary the grouping of students into classes, the assignment of students to teachers, and the amount of learning decay

Leads Test Rejection Rates. 100 Replications. Vertically scaled test scores. λ=1. Correlation of student fixed effect with scorebase is .5.
Row 1: Rejection rate of test with future teachers indicators.
Row 2: Rank Correlations from GRW (2011).

Leads Test λ=1	Cluster at Student Level 4 cohorts			
Estimator Assignment Mechanism	POLS	DOLS	RE	FE
Random Grouping	0.05	0.04	0.05	0.03
Random Assignment	.88	.84	.89	.63
Dynamic Grouping	1	1	1	0.18
Random Assignment	.85	.84	.88	.58
Dynamic Grouping	1	1	1	1
Positive Assignment	.89	.84	.90	-.29
Dynamic Grouping	1	1	1	1
Negative Assignment	.64	.82	.70	.75
Heterog. Grouping	0.94	0.73	0.88	0.07
Random Assignment	.82	.80	.85	.64
Heterog. Grouping	1	1	1	0.09
Positive Assignment	.81	.90	.91	.63
Heterog. Grouping	1	1	1	0.09
Negative Assignment	.39	.43	.55	.63

Leads Test Rejection Rates. 100 Replications. Vertically scaled test scores. λ=1. Correlation of student fixed effect with scorebase is .5. Row 1: Rejection rate of test with future estimated teacher effects. Row 2: Rank Correlations from GRW (forthcoming)

Leads Test λ=1	Cluster at School Level 4 Cohorts				Cluster at School Level 16 Cohorts			
Estimator Assignment Mechanism	POLS	DOLS	RE	FE	POLS	DOLS	RE	FE
Random Grouping	0.15	0.03	0.05	0.07	0.05	0.08	0.07	0.09
Random Assignment	.88	.84	.89	.63	0.97	0.83	0.97	0.69
Dynamic Grouping	0.97	0.02	0.57	0.16	0.51	0.02	0.15	0.05
Random Assignment	.85	.84	.88	.58	0.94	0.83	0.96	0.68
Dynamic Grouping	1	1	1	1	1	1	1	1
Positive Assignment	.89	.84	.90	-.29	0.92	0.81	0.92	-0.36
Dynamic Grouping	0.74	1	0.8	1	0.79	1	0.88	1
Negative Assignment	.64	.82	.70	.75	0.38	0.80	0.38	0.76
Heterog. Grouping	0.7	0.25	0.43	0.1	0.21	0.09	0.14	0.1
Random Assignment	.82	.80	.85	.64	0.94	0.83	0.96	0.69
Heterog. Grouping	1	1	1	0.06	1	1	1	0.06
Positive Assignment	.81	.90	.91	.63	0.93	0.90	0.94	0.69
Heterog. Grouping	0.19	0.51	0.2	0.06	0.19	0.74	0.36	0.09
Negative Assignment	.39	.43	.55	.63	0.46	0.56	0.65	0.69

Concluding Remarks

- New one *df* versions of the Hausman and feedback (falsification) tests are easy to compute and have better properties than the original versions, especially with cohort-level clustering.
- Tests have about the right size under random grouping of students and random assignment of teachers to classrooms.
- For most of the estimators and across most scenarios, the tests badly over reject under random assignment of teachers if the students are grouped nonrandomly. Yet many of the estimators perform very well for estimating teacher VAMs.

- In cases of nonrandom assignment of teachers – static and especially dynamic – the tests detect the nonrandom assignment. But estimators such as DOLS often do well under nonrandom teacher assignment.
- Sometimes the falsification test fails to reject FE when it is easily the worst of the estimators.
- Findings help clear up some puzzles in the literature. Kane and Staiger find VAMs work well, but Rothstein rejects random assignment.

- Strong showing of DOLS for estimating VAMs suggest that viewing the estimation and testing problem from a modern treatment effect perspective is preferred to a more structural approach.
- DOLS is a simple treatment effects estimator when teacher assignment is “unconfounded” conditional on past test scores. Unconfoundedness generally cannot be tested without imposing other assumptions.