Evaluating Teacher Value Added Estimation Methods using Structural Models and Noisy Achievement Data

Brian Stacy   JR Lockwood   Dan McCaffrey

March 8, 2012
Big Picture Questions

Test scores measure a student’s level of knowledge with error

- How can test score measurement error affect teacher value added estimation?
- Are the methods designed to correct measurement error bias worth doing?
Narrow Questions

- When will test measurement error bias estimates?
- What solutions are available? Valid? Practical?
- Does the bias seem to noticeably affect estimates?
- What happens to teacher rankings when you correct for bias?
Test scores are an Imperfect Measure of Achievement

Measurement error can be classified in two categories:

- **Test measurement error**
  - Small number of items on tests
  - Better than average test-taking skills (Tests measure multiple skills. Some of them you don’t want to measure)

- **Short term changes in ability to answer test questions** (Not having to do with knowledge creation or loss):
  - i.e. getting sick, distractions in testing environment
  - Could also consider cramming or cheating as measurement error
Attenuation Bias

- Having an inaccurate measure can create bias
- Let
  \[ A_t^* = \alpha + \beta A_{t-1}^* + u \]
  
  \( A_{t-1} \) is achievement in year \( t \) measured with error.
- Classical assumptions: \( A_{t-1} = A_{t-1}^* + e \)
- 
  \[ \hat{\beta}_{OLS} \sim \frac{\text{Cov}(A_{t-1}, A_t)}{\text{Var}(A_{t-1})} = \beta \left( \frac{\sigma_{A_{t-1}}^2}{\sigma_{A_{t-1}}^* + \sigma_e^2} \right) \]
- As you add more covariates the attenuation gets worse, and bias possibly transmitted to other covariates
1. What solutions are available and are they practical or valid?
   - **Two Models:**
     - First difference IV estimator (IVFD)
     - OLS estimator including lagged scores as regressors (DOLS).
   - **Two Approaches:**
     - Subject scores or lagged scores as IVs
     - Information on variance of measurement error to adjust estimates
   * No estimator fully accounts for all sources of measurement error
   * Find that IVFD estimator is unstable
     - Correcting measurement error might make things worse
2. What does this mean for those estimating teacher value added or other effects?
   * Some differences after correcting for measurement error
   * In the case of DOLS, find that using SEMs to correct measurement error produces similar results using other subjects as IVs
   * Also, adding in classroom aggregate scores produces teacher effect estimates similar to SEM correction and IV correction
Data

- County-wide school district
- Panel data:
  - Student-teacher links
  - Student Characteristics (demographics, special status information)
- Standardized scores because of change in test and lack of vertical scale
Data

Mathematics scores used in analysis

Table: Characteristics of Data Set

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>63,205</td>
</tr>
<tr>
<td>Years</td>
<td>2006-2011</td>
</tr>
<tr>
<td>Grades</td>
<td>3-8</td>
</tr>
<tr>
<td>Cohorts</td>
<td>3</td>
</tr>
<tr>
<td>Teachers</td>
<td>1376</td>
</tr>
</tbody>
</table>
Assume that model is linear, **constant across grades and years**, and effects decay at constant, geometric rate:

\[ A_{it}^* = \alpha_t + \lambda A_{it-1}^* + X_{it}\beta + c_i + \epsilon_{it} \]

Note: This equation uses true achievement, not measured achievement.
Assume that model is linear, **constant across grades and years**, and effects decay at constant, geometric rate:

\[ A_{it}^* = \alpha_t + \lambda A_{it-1}^* + X_{it} \beta + c_i + \epsilon_{it} \]

Note: This equation uses true achievement, not measured achievement

With measured achievement \((A_{it} = A_{it}^* + e_{it})\) we have:

\[ A_{it} = \alpha_t + \lambda A_{it-1} + X_{it} \beta + c_i + \epsilon_{it} + e_{it} - \lambda e_{it-1} \]
When Should You Correct for Measurement Error?

- Typically interested in estimating teacher effects not structural parameter $\lambda$
- Is assignment related more to lagged achievement or lagged test score?
- Test Score:
  - Control for past test score
- Achievement:
  - Control for past achievement
  - Correct for measurement error
- Show that correcting can produce somewhat different rankings
First Differencing the Structural Equation

\[ A_{it} = \alpha_t + \lambda A_{it-1} + X_{it} \beta + c_i + \epsilon_{it} + e_{it} - \lambda e_{it-1} \]

- First difference structural equation to remove student learning heterogeneity (\( \Delta \) denotes differenced variable)

\[ \Delta A_{it} = \Delta \alpha_t + \lambda \Delta A_{it-1} + \Delta X_{it} \beta + \Delta \epsilon_{it} + \Delta e_{it} - \lambda \Delta e_{it-1} \]

- Suppose the structural model is really:

\[ A_{it} = \alpha_t + \lambda_t A_{it-1} + \Delta X_{it} \beta + \Delta \epsilon_{it} + \Delta e_{it} - \lambda \Delta e_{it-1} \]

- First Differencing is messy in this case
Estimation of $\lambda$ Separately by Grade

- Estimates of $\lambda$ done grade by grade vary widely
- Measurement error corrections exacerbate problems

IV: IV regression using lag2 and lag3 reading and ELA, lag3 math
Findings and concern for using IVFD

- Results are very sensitive to the sample used.
- Could be caused by the structural model differencing by grade, year, or cohort
  - Would need to estimate model separately at least by grade and possibly year and cohort
  - With no pooling, can have weak instrument problem
- In any case, difficult to use IVFD estimators
- Not clear correcting for measurement error improves estimates
DOLS Estimation

Going back to the structural model:

\[ A_{it} = \alpha_t + \lambda A_{it-1} + X_{it}\beta + c_i + \epsilon_{it} + e_{it} - \lambda e_{it-1} \]

- Easy to deal with differences across grades
- Can include grade or grade by year interactions
  - Including interactions doesn't significantly affect teacher effect estimates
  - Correlations typically around .99
- \( c_i \) term can be an issue, but may be relatively small with lots of controls
DOLS Measurement Error Corrections

We will explore three of the most common ways of getting around measurement error bias

1. Other subject and further lag scores as IVs
   - Scores used as IVs also measured with error
   - IVs only valid if error in other scores independent
   - Not likely true (e.g. getting sick affects math and ELA scores, good test taking skills leads to serial correlation)
   - Perhaps can partially fix attenuation
2. Use known measurement error variances to fix regression cross product ($X'X$) matrix
  - Subtract measurement error variance from matrix element corresponding to lag test score(s)
  - Procedure can be performed by using eivreg command in stata
DOLS Measurement Error Corrections

3. Regression Calibration using known measurement error variances
   ▶ OLS estimation: OLS regression of $A_{it}$ on $A_{it-1}$ and $X_{it}$
   ▶ Regression Calibration: OLS regression of $A_{it}$ on $\hat{A}_{it-1}$ and $X_{it}$ where $\hat{A}_{t-1}$ is the best linear predictor of $A_{it-1}$ given $A_{it-1}$ and $X_{it}$.
   ▶ Valid under classical assumptions

- A key feature of method 2 and 3 is that measurement error variance used is correct
  ▶ Typically based on SEMs
  ▶ Don’t pick up variance due to getting sick or other day to day changes in ability to answer questions

- Procedures don’t produce consistent estimates with serial correlation
Estimation of $\lambda$ Separately by Grade

- Estimates of $\lambda$ still vary widely

IV: IV regression using lag1 and lag2 reading and ELA, lag2 math
Teacher Effect Correlations across methods

Table: Correlation table for teacher effect estimates

<table>
<thead>
<tr>
<th>Variables</th>
<th>OLS</th>
<th>IV</th>
<th>Eivreg</th>
<th>Regression Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>0.778</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eivreg</td>
<td>0.844</td>
<td>0.993</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Regression Calibration</td>
<td>0.844</td>
<td>0.993</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Equation:

\[ A_{it} = \alpha_t + \lambda A_{it-1} + X_{it} \beta + c_i + \eta_{it} \]

OLS: OLS regression of parameters  
IV: IV regression using lag1 and lag2 reading and ELA, lag2 math  
Eivreg: Adjust lag1 math score element in X’X matrix using SEMs  
Regression Calibration: Form BLUP for lag1 math score using SEMs
Conclusions

- Fixing measurement error tricky problem
- Measurement error large enough to alter estimates
- First difference estimators are unstable
  - Even more unstable using estimators that correct for measurement error
- IV and strategies that correct estimates using known measurement error variances give similar estimates
- Controlling for class aggregates produces similar teacher rankings estimators that explicitly correct for measurement error
Including Classroom Aggregates

- Kane and Staiger (2008) and Chetty, Rockoff, Friedman (2012) estimate with lagged scores as well as classroom aggregates.
- Find no bias in estimates.
- How could this be given measurement error?
- How do measurement error corrected estimates compare?
Teacher Effect Correlations

- Adding in classroom aggregates, such as lag score, gives similar teacher rankings as the estimators that correct for measurement error.
- This could be due to the class aggregates proxying for that student’s own ability.

Table: Correlation table for teacher effect estimates

<table>
<thead>
<tr>
<th>Variables</th>
<th>OLS</th>
<th>IV</th>
<th>Eivreg</th>
<th>Aggregates</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>0.778</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eivreg</td>
<td>0.844</td>
<td>0.993</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Aggregates</td>
<td>0.766</td>
<td>0.956</td>
<td>0.954</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Equation:

\[ A_{it} = \alpha_t + \lambda A_{it-1} + X_{it}\beta + c_i + \eta_{it} \]

Aggregates: Same as Naive, but includes class average lag score, English proficiency, gifted, special ed, frl, and race. Also uses average residuals instead of teacher dummies.
Contrasting Methods

- Including classroom aggregates typically means computing average residuals
  - Assumes teacher assignment unrelated to student controls
  - Estimates could suffer if not true
- IV methods require multiple lags or other subjects
  - Imposes data requirements
  - Also assumes IVs are valid
- Using SEMs to correct for measurement error requires additional information also
  - Requires accurate estimate of total variance of measurement error
  - SEMs aren’t reflective of all forms of measurement error
The point \((4, k)\) is a solution to the equation \(3x + 2y = 12\). What is the value of \(k\)?

A. -3
B. 0
C. 2
D. 3
E. 4

Source: 2011 Mathematics NAEP exam, grade 8
Testing the Structural Equation

- Simple test of the structural model

\[ \Delta A_t = \Delta \alpha_t + \lambda \Delta A_{t-1} + \Delta X_t \beta + \Delta \epsilon_t + \Delta e_t - \lambda \Delta e_{t-1} \]

- Test model by interacting \( \Delta A_{t-1} \) with grade
- Should see no differences in estimates of \( \lambda \) across grades

Use Math lag2 as instrument for lag score gain\(^1\)

<table>
<thead>
<tr>
<th>Estimates for lag score gain interacted with grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
</tr>
<tr>
<td>Grade 5</td>
</tr>
<tr>
<td>Grade 6</td>
</tr>
<tr>
<td>Grade 7</td>
</tr>
<tr>
<td>Grade 8</td>
</tr>
</tbody>
</table>

Joint test of equal coefficients: \( F(4, 13024) = 40.07 \) p-value=0.000

\(^1\)Also performed test using other subject scores and further lags as IVs and also reject
Estimates of $\lambda$

Given that coefficient on lag score necessarily affected by measurement error, interesting to compare estimates across estimators

<table>
<thead>
<tr>
<th></th>
<th>Estimates of $\lambda$</th>
<th>Obs= 60557</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>.618 (.003)</td>
<td>60557</td>
</tr>
<tr>
<td>IV</td>
<td>.940 (.005)</td>
<td></td>
</tr>
<tr>
<td>Eivreg</td>
<td>.894 (.004)</td>
<td></td>
</tr>
</tbody>
</table>

Equation:

$A_{it} = \alpha_t + \lambda A_{it-1} + X_{it}\beta + c_i + \eta_{it}$

OLS: OLS regression of parameters
IV: IV regression using lag1 and lag2 reading and ELA, lag2 math
Eivreg: Adjust lag1 math score element in $X'X$ matrix using SEMs
Regression Calibration estimates differ from Eivreg at 7th decimal place