Teacher Effects on Student Achievement and Height: A Cautionary Tale

Marianne Bitler\textsuperscript{1}
Sean Corcoran\textsuperscript{2}
Thurston Domina\textsuperscript{1}
Emily Penner\textsuperscript{1}

\textsuperscript{1}University of California—Irvine
\textsuperscript{2}NYU Steinhardt School of Culture, Education, & Human Development

Thanks to NICHD for financial support #1PO1HD065704-01A1

June 11, 2015
Estimates of teacher value-added

Data linking students to teachers has made it possible to estimate the contribution teachers make to student achievement. These estimates are called “teacher effects” or “value-added” measures (VAMs)—the extent to which the achievement of teacher j’s students differs, on average, from that predicted by their past achievement and other covariates:

\[ Y_{it} = \alpha Y_{it-1} + X_{it}' \beta + u_j + e_{it} \]

- \( Y_{it} \) and \( Y_{it-1} \) = current and lagged test score
- \( X_{it} \) = student and other covariates
- \( u_j \) = teacher effect
Estimates of teacher value-added

Uses of teacher value-added

VAM estimates are used for a variety of purposes, including quantifying the overall importance of teachers to student achievement. Teacher effects on short-run achievement are large, and these effects are correlated with long-run outcomes, including earnings (Chetty et al., 2014).

Teacher effect size estimates: $1\sigma \rightarrow$

- Rivkin et al. 2005 (0.10$\sigma$ reading, 0.11$\sigma$ math)
- Rockoff 2004 (0.10$\sigma$ R, 0.11$\sigma$ M)
- Kane & Staiger 2008 (0.18$\sigma$ R, 0.22$\sigma$ M)
- Buddin 2010 (0.19$\sigma$ R, 0.28$\sigma$ M)
- Papay 2011 (0.02$\sigma$ - 0.16$\sigma$ various)
- Corcoran & Jennings 2011 (0.16$\sigma$ - 0.26$\sigma$ various)
Uses of teacher value-added

VAMs are increasingly being used by states and districts to identify high- and low-performing teachers.

- Many teacher evaluation systems now use value-added measures as significant criteria in promotion and dismissal.
- 16 states + D.C. require 50% or more of teachers annual evaluations to be based on VAM or comparable growth measures: AK, CO, DC, FL, GA, HI, IL, LA, MI, MS, NV, NM, NY, OH, OK, PA, TN.
- VAMs are sometimes used to award bonuses and determine compensation.
- In a few cases, VAMs have been publicly reported in the media.
Uses of teacher value-added

VAMs are increasingly being used by states and districts to identify high- and low-performing teachers.

- Many teacher evaluation systems now use value-added measures as significant criteria in promotion and dismissal.
- 16 states + D.C. require 50% or more of teachers annual evaluations to be based on VAM or comparable growth measures: AK, CO, DC, FL, GA, HI, IL, LA, MI, MS, NV, NM, NY, OH, OK, PA, TN.
- VAMs are sometimes used to award bonuses and determine compensation.
- In a few cases, VAMs have been publicly reported in the media.
- Typically, categorical ratings are assigned to teachers based on their position in the distribution of VAMs:
Uses of teacher value-added

**Figure 5. Determining Teacher Growth Ratings**

- **MGP**
  - Greater than 1.5 SD* above mean (MGP > 88)
  - Between 1 SD below mean to 1.5 SD above mean (MGP between 41-88)
  - 1.5 to 1 SD below mean (MGP between 35-41)
  - More than 1.5 SD below mean (MGP < 35)

- **Confidence Range**
  - Lower Limit > Mean (Lower Limit > 52)
  - Upper Limit < Mean (Upper Limit < 52)
  - Upper Limit (< .75 SD below mean) (Upper Limit < 43)
  - Any

- **Growth Rating**
  - Highly Effective: Well above state average for similar students (18—20 points)
  - Effective: Equal to state average for similar students (9—17 points)
  - Developing: Below average for similar students (3—6 points)
  - Ineffective: Well below average for similar students (0—2 points)

*Standard deviation
Concerns raised about VAMs

The high-stakes use of value-added to evaluate teachers has been controversial, with concerns raised about:

- **Bias:** teacher effects may reflect omitted variables and/or selection on unobservables (e.g., Rothstein, 2010; Horvath, 2015)
- **Measurement error:** teacher effect estimates are noisy and do not consistently rank teachers across years or subjects (e.g., McCaffrey et al., 2009; Schochet & Chiang, 2013; Papay, 2011).

Counterargument: VAMs are related to future outcomes, and are better than existing measures of teacher quality or subjective ratings (Glazerman 2010, 2011; Kane & Staiger, 2008; Kane et al., 2013).
What we do

Using data from NYC, we apply traditionally-estimated VAM models to estimate teacher “effects” on height. Why do this?

- Potential falsification test: teachers—at least in the U.S.—should not have a causal effect on height.
- Height is distributed normally in the population, should be measured with less error than achievement, and should be free of peer effects. There are few other student-level outcomes to which one could apply this approach.
- Results could be informative about the properties of VAM models, the importance of sorting, and noise.
What we do—and preview

We find significant “effects” of teachers on height, and consider three possible interpretations:

1. Sorting on factors related to height that are also related to achievement. This could mean achievement VAMs are biased.
What we do—and preview

We find significant “effects” of teachers on height, and consider three possible interpretations:

1. Sorting on factors related to height that are also related to achievement. This could mean achievement VAMs are biased.

2. Effects are spurious variation, or random “noise.” Differences attributed to teachers are simply idiosyncratic variation across relatively small samples.
What we do—and preview

We find significant “effects” of teachers on height, and consider three possible interpretations:

1. Sorting on factors related to height that are also related to achievement. This could mean achievement VAMs are biased.

2. Effects are spurious variation, or random “noise.” Differences attributed to teachers are simply idiosyncratic variation across relatively small samples.

3. Sorting on factors related to height that are uncorrelated with achievement. This type of sorting would be less worrisome.
What we do—and preview

How we evaluate these explanations:

- Sorting on height
  - Look at correlation of VAMs on height and achievement
  - Look at systematic sorting to classrooms and teachers on lagged height and achievement
What we do—and preview

How we evaluate these explanations:

- Sorting on height
  - Look at correlation of VAMs on height and achievement
  - Look at systematic sorting to classrooms and teachers on lagged height and achievement

- The role of noise
  - Look at covariance in teacher effects across years for teachers with multiple years of classroom data
  - Estimate 3-level models (teacher, classroom, student)
  - Random permutation tests
Data

We use a panel of students grades 4-5 in NYC public schools between 2007 and 2010.

- Students are linked to math and ELA teachers, and to annual “Fitnessgram” results.
- A large number of students and teachers, and (in some cases) many students per teacher (as many as four cohorts).
- Teacher links are available for grades 6-8, but we focus on students in self-contained classrooms.
- Covariates include age, gender, race/ethnicity, ELL, special education.
Data

We have estimated similar models using the ECLS-K (not included in this presentation). The ECLS-K has advantages and disadvantages:

- It is a national study in which trained assessors measured participants’ height in both the fall and spring of their kindergarten year.
- Within-school sorting is probably minimal in kindergarten.
- Achievement and height are more finely measured in the ECLS-K.
- But fewer students per teacher, and teachers are observed with only one cohort.
Fitnessgram

- Administered since 2005-06
- Conducted by trained personnel (usually PE teacher) using a common procedure and recommended digital scale.
NYC data

Additional data details:

- ELA and math scores are standardized by subject, grade, and year.
- Height is standardized by grade and year, with outliers dropped ($\geq 4\sigma$ from the age-gender mean).
- We alternatively standardized height by gender and age in months (produced similar results).
- Students included in teacher effect models are required to have lagged values of the dependent variable, and the teacher must have at least seven students with the necessary data.
NYC data—student height measures

Bitler et al. (2015)
NYC data—student math scores
NYC data—students per teacher

[Histograms showing the distribution of students per teacher for grades 4 and 5 in math and height.]
## NYC data—teachers and students

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Math</th>
<th>ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 4</td>
<td>Grade 5</td>
<td>Grade 4</td>
</tr>
<tr>
<td>Unique teachers (N)</td>
<td>4,263</td>
<td>3,687</td>
<td>4,721</td>
</tr>
<tr>
<td>Mean years observed</td>
<td>1.90</td>
<td>1.98</td>
<td>1.88</td>
</tr>
<tr>
<td>Students per teacher:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>36.0</td>
<td>39.0</td>
<td>38.7</td>
</tr>
<tr>
<td>SD</td>
<td>22.9</td>
<td>25.5</td>
<td>24.5</td>
</tr>
<tr>
<td>p25</td>
<td>19</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>p50</td>
<td>27</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>p90</td>
<td>71</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>Unique classrooms (N)</td>
<td>7,594</td>
<td>6,848</td>
<td>8,712</td>
</tr>
<tr>
<td>Students per classroom:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>20.0</td>
<td>20.8</td>
<td>20.9</td>
</tr>
<tr>
<td>SD</td>
<td>5.4</td>
<td>6.4</td>
<td>5.1</td>
</tr>
<tr>
<td>p25</td>
<td>17</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>p50</td>
<td>20</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>p90</td>
<td>26</td>
<td>28</td>
<td>27</td>
</tr>
</tbody>
</table>
NYC data—student means, grades 4-5

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All linked obs</td>
<td>Height sample</td>
</tr>
<tr>
<td>ELA z-score</td>
<td>0.027</td>
<td>0.071</td>
</tr>
<tr>
<td>Math z-score</td>
<td>0.033</td>
<td>0.102</td>
</tr>
<tr>
<td>Height (inches)</td>
<td>54.662</td>
<td>54.587</td>
</tr>
<tr>
<td>Height z-score</td>
<td>-0.032</td>
<td>-0.043</td>
</tr>
<tr>
<td>Female</td>
<td>0.506</td>
<td>0.509</td>
</tr>
<tr>
<td>White</td>
<td>0.156</td>
<td>0.169</td>
</tr>
<tr>
<td>Black</td>
<td>0.283</td>
<td>0.275</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.392</td>
<td>0.376</td>
</tr>
<tr>
<td>Asian</td>
<td>0.162</td>
<td>0.181</td>
</tr>
<tr>
<td>Low income</td>
<td>0.798</td>
<td>0.804</td>
</tr>
<tr>
<td>LEP</td>
<td>0.119</td>
<td>0.102</td>
</tr>
<tr>
<td>Special ed</td>
<td>0.119</td>
<td>0.115</td>
</tr>
<tr>
<td>English at home</td>
<td>0.585</td>
<td>0.576</td>
</tr>
<tr>
<td>Recent immigrant</td>
<td>0.130</td>
<td>0.117</td>
</tr>
<tr>
<td>Same math/ELA teacher</td>
<td>0.900</td>
<td>0.883</td>
</tr>
<tr>
<td>Manhattan</td>
<td>0.133</td>
<td>0.119</td>
</tr>
<tr>
<td>Bronx</td>
<td>0.207</td>
<td>0.165</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>0.312</td>
<td>0.340</td>
</tr>
<tr>
<td>Queens</td>
<td>0.284</td>
<td>0.302</td>
</tr>
<tr>
<td>2007</td>
<td>0.241</td>
<td>0.167</td>
</tr>
<tr>
<td>2008</td>
<td>0.243</td>
<td>0.225</td>
</tr>
<tr>
<td>2009</td>
<td>0.251</td>
<td>0.274</td>
</tr>
<tr>
<td>2010</td>
<td>0.264</td>
<td>0.334</td>
</tr>
</tbody>
</table>

N 239,577 153,297 182,623 236,983 143,774 180,637
## NYC data—student-level correlations

<table>
<thead>
<tr>
<th>Correlations between:</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math and ELA</td>
<td>0.688***</td>
<td>0.585***</td>
</tr>
<tr>
<td>Math and height</td>
<td>−0.059</td>
<td>−0.068***</td>
</tr>
<tr>
<td>ELA and height</td>
<td>−0.046***</td>
<td>−0.042***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation with lag:</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>0.701***</td>
<td>0.757***</td>
</tr>
<tr>
<td>ELA</td>
<td>0.683***</td>
<td>0.646***</td>
</tr>
<tr>
<td>Height</td>
<td>0.799***</td>
<td>0.793***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlations between changes in:</th>
<th>Grade 4</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math and ELA</td>
<td>0.158***</td>
<td>0.140***</td>
</tr>
<tr>
<td>Math and height</td>
<td>0.002</td>
<td>0.007**</td>
</tr>
<tr>
<td>ELA and height</td>
<td>0.013***</td>
<td>−0.006*</td>
</tr>
</tbody>
</table>
Baseline value-added model specifications

Basic model:

\[ Y_{it} = \alpha Y_{it-1} + X'_{it} \beta + \gamma_t + u_j + e_{it} \]

- The \( u_j \) are often assumed to be random effects, estimated using shrinkage or Empirical Bayes estimators, or fixed effects.
- Use BLUPs post-estimation, and mean residuals scaled by a shrinkage factor (Kane, Staiger, & Rockoff, 2008).
- The variance components \( \sigma_u \) and \( \sigma_e \) are estimated parameters.

\[ \lambda_j = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_j} \]
VAM model specifications

We estimate the teacher effects under the random effects assumption (using BLUPs and mean residuals approach) and under a fixed effects assumption (also adjusting the estimated $u_j$ by the shrinkage factor). Covariates $X_{it}$ include:

- Three way interaction: gender, race, and age
- Recent immigrant, LEP, English at home, special education, low income, borough of residence
- Height models add days between measurements
VAM model specifications

We estimate the teacher effects under the random effects assumption (using BLUPs and mean residuals approach) and under a fixed effects assumption (also adjusting the estimated $u_j$ by the shrinkage factor). Covariates $X_{it}$ include:

- Three way interaction: gender, race, and age
- Recent immigrant, LEP, English at home, special education, low income, borough of residence
- Height models add days between measurements

As others do, we find strong correlations at the teacher level between RE and FE estimates (0.71 to 0.96, depending on the grade and measure).
VAM model specifications

We also estimate versions with school fixed effects ($\phi_s$):

$$Y_{it} = \alpha Y_{it-1} + X_{it}' \beta + \gamma_t + \phi_s + u_j + e_{it}$$

Models with school effects are more common in research than in practical applications. In our context we were concerned that variability in height could be driven by school-level factors.
SD of estimated teacher effects–grade 4

<table>
<thead>
<tr>
<th>Model:</th>
<th>Height</th>
<th>Math</th>
<th>ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Baseline models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.218</td>
<td>0.286</td>
<td>0.256</td>
</tr>
<tr>
<td>FE (adj.)</td>
<td>0.250</td>
<td>0.344</td>
<td>0.278</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.169</td>
<td>0.216</td>
<td>0.184</td>
</tr>
<tr>
<td>FE w/school effects (adj.)</td>
<td>0.166</td>
<td>0.202</td>
<td>0.172</td>
</tr>
<tr>
<td>N of teachers</td>
<td>4,262</td>
<td>4,721</td>
<td>4,366</td>
</tr>
<tr>
<td>Mean students per teacher</td>
<td>36.0</td>
<td>38.7</td>
<td>35.9</td>
</tr>
</tbody>
</table>

*Bitler et al. (2015)*

Teacher effects on height
SD of estimated teacher effects–grade 5

<table>
<thead>
<tr>
<th>Model</th>
<th>Height</th>
<th>Math</th>
<th>ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Baseline models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.210</td>
<td>0.253</td>
<td>0.210</td>
</tr>
<tr>
<td>FE (adj.)</td>
<td>0.315</td>
<td>0.258</td>
<td>0.240</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.157</td>
<td>0.199</td>
<td>0.155</td>
</tr>
<tr>
<td>FE w/school effects (adj.)</td>
<td>0.160</td>
<td>0.189</td>
<td>0.145</td>
</tr>
<tr>
<td>N of teachers</td>
<td>3,687</td>
<td>4,249</td>
<td>3,978</td>
</tr>
<tr>
<td>Mean students per teacher</td>
<td>39.0</td>
<td>42.5</td>
<td>39.5</td>
</tr>
</tbody>
</table>
Teacher effects on height

To put in perspective: a $0.22\sigma$ increase in height amounts to:

- 0.68-inch gain in stature for 4th graders
- 0.72-inch gain in stature for 5th graders
- (Roughly $1/3\sigma$ in year-to-year growth)
Distribution of estimated teacher effects—height

```
Grade 4

Random effects: height

Grade 5

Fixed effects: height
```
Distribution of estimated teacher effects—math

![Graphs showing distribution of teacher effects for Grades 4 and 5 for both random and fixed effects in math.](image)
Do teacher effects on height reflect bias?

Is there systematic sorting on height—or factors related to height—that could potentially bias achievement VAMs?
Do teacher effects on height reflect bias?

Is there systematic sorting on height—or factors related to height—that could potentially bias achievement VAMs?

- Correlate teacher effects on height and achievement
- Examine systematic sorting on lagged height (Horvath, 2015)
# Pairwise correlations in teacher effects—grade 4

<table>
<thead>
<tr>
<th>Grade 4</th>
<th>RE</th>
<th>FE (adj)</th>
<th>RE w/ school effects</th>
<th>FE w/ school effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height VAM:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>-0.019</td>
<td>-0.014</td>
<td>-0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>FE (adj)</td>
<td>-0.030*</td>
<td>0.199*</td>
<td>-0.022</td>
<td>-0.023</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.000</td>
<td>-0.003</td>
<td><strong>0.002</strong></td>
<td>0.002</td>
</tr>
<tr>
<td>FE w/school effects</td>
<td>-0.002</td>
<td>-0.004</td>
<td>0.001</td>
<td><strong>0.000</strong></td>
</tr>
</tbody>
</table>

| ELA VAM: | | | | |
| RE | **0.697*** | 0.597* | 0.521* | **0.519*** |
| FE (adj) | 0.658* | **0.689*** | 0.477* | **0.475*** |
| RE w/school effects | 0.525* | 0.432* | **0.646*** | 0.643* |
| FE w/school effects | 0.522* | 0.428* | 0.643* | **0.641*** |
### Pairwise correlations in teacher effects—grade 5

<table>
<thead>
<tr>
<th></th>
<th>Math VAM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 5</td>
<td>RE</td>
</tr>
<tr>
<td><strong>Height VAM:</strong></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.016</td>
</tr>
<tr>
<td>FE (adj)</td>
<td>0.009</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.001</td>
</tr>
<tr>
<td>FE w/school effects</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>ELA VAM:</strong></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td><strong>0.557</strong></td>
</tr>
<tr>
<td>FE (adj)</td>
<td><strong>0.511</strong></td>
</tr>
<tr>
<td>RE w/school effects</td>
<td><em>0.425</em>*</td>
</tr>
<tr>
<td>FE w/school effects</td>
<td><em>0.424</em>*</td>
</tr>
</tbody>
</table>
Horvath (2015) identified schools that practice nonrandom classroom assignment by testing for systematic variation in lagged student characteristics across classrooms within schools, grades, and years. For example:

$$Y_{it-1} = u_c + \phi_{sgt} + w_{it}$$

For each school test the null hypothesis that the classroom effects are zero. (A p-value less than 0.05 suggests schools “track” students to classrooms). She performed a similar test for teacher “matching,” defined as persistent tracking to specific teachers.
Tests for tracking on lagged student characteristics
Tests for tracking on lagged student characteristics

Summary:

- Math: 64.6% (62.6%) of schools track in 4th (5th) grade—compare to Horvath’s 60% for North Carolina
- Height: 10.1% (11.2%) of schools track in 4th (5th) grade
- It is common for a school to track in both 4th and 5th grade in math, but rare for height
Are teacher effects on height just noise?

Is there a *persistent* teacher effect on height, or are they spurious?
Are teacher effects on height just noise?

Is there a persistent teacher effect on height, or are they spurious?

- Correlate teacher effects across years, for those with multiple years of classroom data
- Estimate 3-level model, allowing for unobserved group-level variability within teacher over time ($u_{jt} = u_j + v_{jt}$)
- $\sigma_u^2$ estimated using covariance in $u_{jt}$ (Kane, Staiger, & Rockoff, 2008). Shrinkage factor:

$$\lambda_j = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_e^2/n_j}$$

- Permutation tests randomly allocating student data to teachers
### Between-year correlations in teacher effects

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>N(4)</th>
<th>N(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Height:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>-0.166</td>
<td>-0.167</td>
<td>3,319</td>
<td>3,135</td>
</tr>
<tr>
<td>FE (adj)</td>
<td>0.001</td>
<td>-0.094</td>
<td>3,319</td>
<td>3,135</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>-0.004</td>
<td>0.007</td>
<td>3,285</td>
<td>3,100</td>
</tr>
<tr>
<td>FE w/school effects (adj)</td>
<td>0.000</td>
<td>0.011</td>
<td>3,285</td>
<td>3,100</td>
</tr>
<tr>
<td><strong>Math:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.557</td>
<td>0.479</td>
<td>4,001</td>
<td>3,885</td>
</tr>
<tr>
<td>FE (adj)</td>
<td>0.587</td>
<td>0.498</td>
<td>4,001</td>
<td>3,885</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.463</td>
<td>0.435</td>
<td>3,988</td>
<td>3,868</td>
</tr>
<tr>
<td>FE w/school effects (adj)</td>
<td>0.471</td>
<td>0.438</td>
<td>3,988</td>
<td>3,868</td>
</tr>
<tr>
<td><strong>ELA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.456</td>
<td>0.408</td>
<td>3,428</td>
<td>3,357</td>
</tr>
<tr>
<td>FE (adj)</td>
<td>0.501</td>
<td>0.453</td>
<td>3,428</td>
<td>3,357</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.247</td>
<td>0.210</td>
<td>3,410</td>
<td>3,345</td>
</tr>
<tr>
<td>FE w/school effects (adj)</td>
<td>0.249</td>
<td>0.214</td>
<td>3,410</td>
<td>3,345</td>
</tr>
</tbody>
</table>
## SD of estimated teacher effects–3-level models

<table>
<thead>
<tr>
<th>Model</th>
<th>Grade 4</th>
<th></th>
<th>Grade 5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Height</td>
<td>Math</td>
<td>ELA</td>
<td>Height</td>
</tr>
<tr>
<td><strong>B. 3-level models (KS&amp;R)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.000</td>
<td>0.163</td>
<td>0.104</td>
<td>0.000</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.000</td>
<td>0.107</td>
<td>0.077</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>C. 3-level models (MLE)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td>0.000</td>
<td>0.199</td>
<td>0.159</td>
<td>0.000</td>
</tr>
<tr>
<td>RE w/school effects</td>
<td>0.000</td>
<td>0.108</td>
<td>0.070</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Bitler et al. (2015)

Teacher effects on height

June 11, 2015
Permutation tests

Impose the null hypothesis of no sorting, no true effects, no peer effects, no systematic measurement error, etc., by randomly allocating students to teachers in our data set.

- Estimate the same models under 499 random permutations of students to teachers (within year), preserving the number of students assigned to each teacher.
- Fully randomized across teachers, and randomized within schools.
- Save teacher effects and estimated standard deviation of teacher effects ($\hat{\sigma}_u$) on each iteration.
Permutation tests

Impose the null hypothesis of no sorting, no true effects, no peer effects, no systematic measurement error, etc., by randomly allocating students to teachers in our data set.

- Estimate the same models under 499 random permutations of students to teachers (within year), preserving the number of students assigned to each teacher.
- Fully randomized across teachers, and randomized within schools.
- Save teacher effects and estimated standard deviation of teacher effects ($\hat{\sigma}_u$) on each iteration.

The distribution of these will be informative about the null: what we would expect to see if there were no effects of any kind.
$\sigma_u$ from permutations—height and math

95th percentiles: 0.058 and 0.056 (height), 0.065 and 0.058 (math)
\( \sigma_u \) from permutations—within school

95th percentiles: 0.134 and 0.126 (height), 0.086 and 0.086 (math)
Discussion—bad news and good news

- **Bad news**
  - Teacher effects appear substantial on an outcome that teachers cannot plausibly affect. In less obvious applications, an analyst might be tempted to interpret these as meaningful (perhaps causal) differences.
Discussion—bad news and good news

- **Bad news**
  - Teacher effects appear substantial on an outcome that teachers cannot plausibly affect. In less obvious applications, an analyst might be tempted to interpret these as meaningful (perhaps causal) differences.

- **Good news**
  - There is little evidence here to suggest that teacher effects on height reflect any kind of unobserved sorting process that might bias VAM estimates of achievement.
  - Due diligence and validation—as being done with achievement VAMs—would prevent the inappropriate use of measures like these which contain no signal.
Discussion—bad news and good news

- **Bad news**
  - Teacher effects appear substantial on an outcome that teachers cannot plausibly affect. In less obvious applications, an analyst might be tempted to interpret these as meaningful (perhaps causal) differences.

- **Good news**
  - There is little evidence here to suggest that teacher effects on height reflect any kind of unobserved sorting process that might bias VAM estimates of achievement.
  - Due diligence and validation—as being done with achievement VAMs—would prevent the inappropriate use of measures like these which contain no signal.

- **Bad news**
  - VAMs appear to contain a lot of noise. Most applications are less obvious than this, and separating the signal from the noise in individual teacher effect estimates is not straightforward.
  - Getting the shrinkage factor “right” may have limited value in purging noise from individual estimates, since it has only modest effect on the relative rankings of teachers.