Teacher Effects on Student Achievement and Height: A Cautionary Tale

Marianne Bitler[1]
Sean Corcoran[2]
Thurston Domina[1]
Emily Penner[1]

[1]University of California—Irvine
[2]NYU Steinhardt School of Culture, Education, & Human Development

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

June 11, 2015

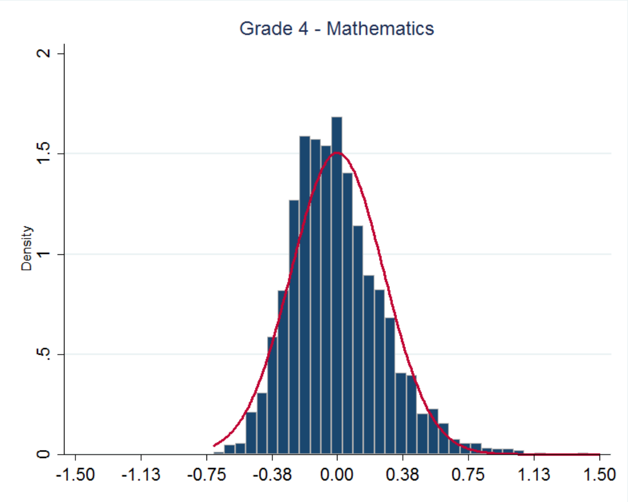
Estimates of teacher value-added

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



Example: NYC 4th grade mathematics, 2007-2010.

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 (0.10σ reading, 0.11σ math)
- Rockoff 2004 (0.10σ R, 0.11σ M)
- Kane & Staiger 2008 (0.18σ R, 0.22σ M)
- Buddin 2010 (0.19σ R, 0.28σ M)
- Papay 2011 (0.02σ - 0.16σ various)
- Corcoran & Jennings 2011 (0.16σ - 0.26σ 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.

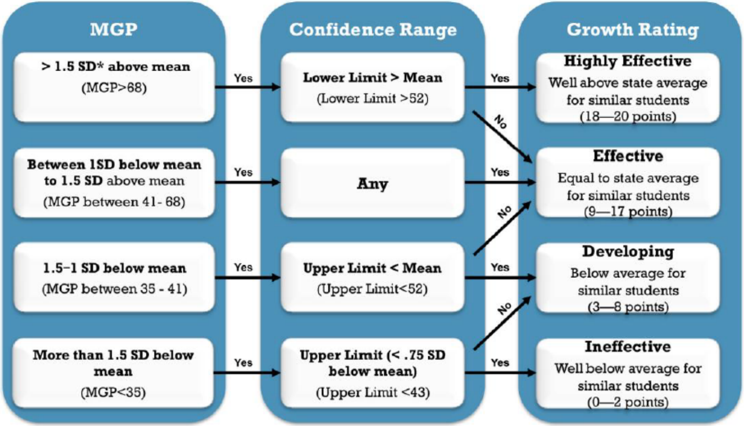
Uses of teacher value-added

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 GrowthRatings



*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

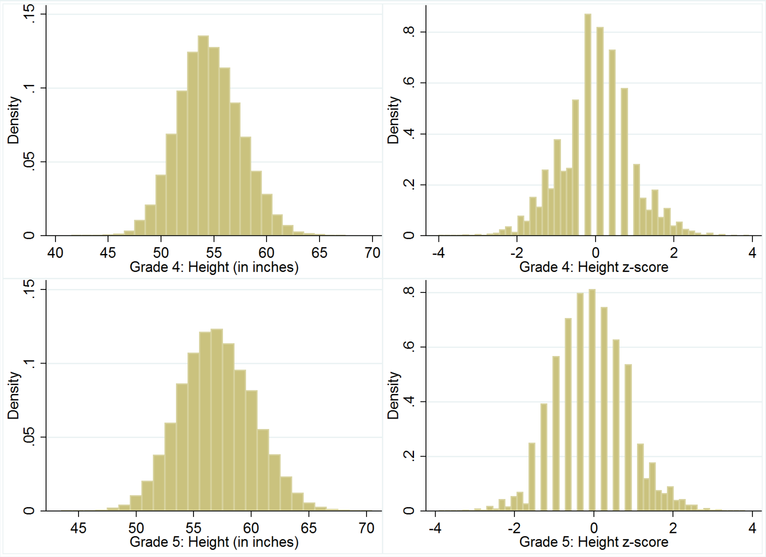
- Administered since 2005-06
- Conducted by trained personnel (usually PE teacher) using a common procedure and recommended digital scale.

NYC data

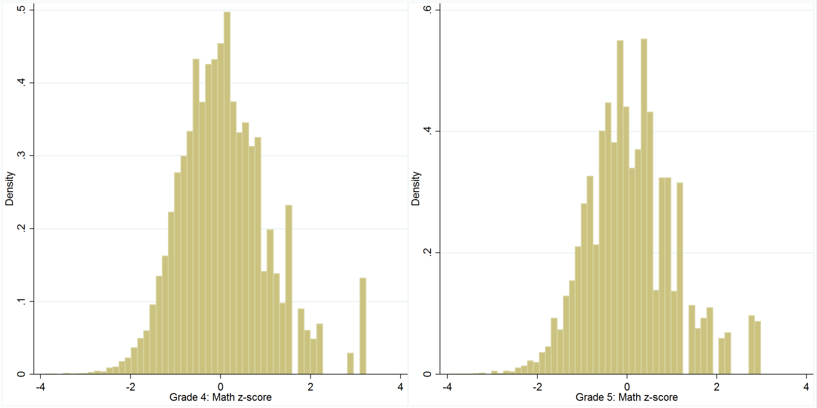
Additional data details:

- 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.

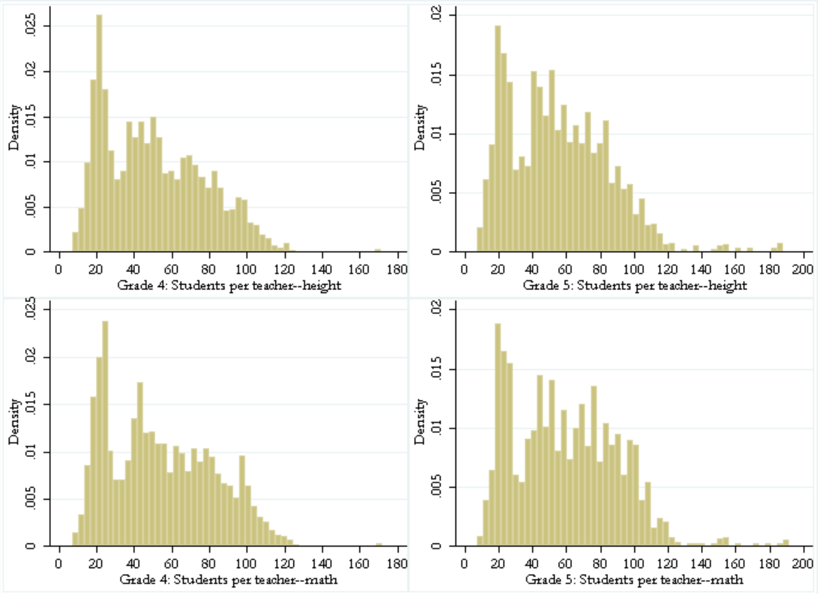
NYC data—student height measures



NYC data—student math scores



NYC data—students per teacher



NYC data—teachers and students

	Height		Math		ELA	
	Grade 4	Grade 5	Grade 4	Grade 5	Grade 4	Grade 5
Unique teachers (N)	4,263	3,687	4,721	4,249	4,366	3,978
Mean years observed	1.90	1.98	1.88	1.94	1.82	1.87
Students per teacher:						
Mean	36.0	39.0	38.7	42.5	35.9	39.5
SD	22.9	25.5	24.5	27.4	22.8	24.9
p25	19	20	20	21	19	20
p50	27	29	29	33	26	29
p90	71	76	77	84	72	78
Unique classrooms (N)	7,594	6,848	8,712	8,138	7,941	7,451
Students per classroom:						
Mean	20.0	20.8	20.9	22.2	19.7	21.1
SD	5.4	6.4	5.1	6.4	5.6	6.3
p25	17	17	18	19	16	18
p50	20	21	21	22	20	21
p90	26	28	27	28	26	28

NYC data—student means, grades 4-5

	Grade 4			Grade 5		
	All linked obs	Height sample	Math sample	All linked obs	Height sample	Math sample
ELA z-score	0.027	0.071	0.051	0.025	0.069	0.047
Math z-score	0.033	0.102	0.079	0.035	0.119	0.087
Height (inches)	54.662	54.587	54.649	57.082	57.003	57.080
Height z-score	-0.032	-0.043	-0.033	-0.035	-0.043	-0.032
Female	0.506	0.509	0.507	0.505	0.507	0.507
White	0.156	0.169	0.162	0.152	0.167	0.157
Black	0.283	0.275	0.281	0.286	0.277	0.283
Hispanic	0.392	0.376	0.386	0.395	0.376	0.388
Asian	0.162	0.181	0.171	0.162	0.181	0.171
Age	9.645	9.626	9.640	10.670	10.647	10.665
Low income	0.798	0.804	0.804	0.799	0.805	0.806
LEP	0.119	0.102	0.105	0.101	0.082	0.086
Special ed	0.119	0.115	0.118	0.116	0.111	0.114
English at home	0.585	0.576	0.582	0.573	0.564	0.570
Recent immigrant	0.130	0.117	0.117	0.148	0.137	0.137
Same math/ELA teacher	0.900	0.883	0.893	0.862	0.858	0.867
Manhattan	0.133	0.119	0.125	0.131	0.115	0.125
Bronx	0.207	0.165	0.184	0.209	0.158	0.181
Brooklyn	0.312	0.340	0.325	0.310	0.348	0.328
Queens	0.284	0.302	0.299	0.287	0.304	0.299
2007	0.241	0.167	0.204	0.247	0.174	0.207
2008	0.243	0.225	0.241	0.244	0.236	0.242
2009	0.251	0.274	0.259	0.254	0.279	0.261
2010	0.264	0.334	0.297	0.255	0.311	0.290
N	239,577	153,297	182,623	236,983	143,774	180,637

NYC data—student-level correlations

Correlations between:	Grade 4	Grade 5
Math and ELA	0.688***	0.585***
Math and height	−0.059	−0.068***
ELA and height	−0.046***	−0.042***

Correlation with lag:	Grade 4	Grade 5
Math	0.701***	0.757***
ELA	0.683***	0.646***
Height	0.799***	0.793***

Correlations between changes in:	Grade 4	Grade 5
Math and ELA	0.158***	0.140***
Math and height	0.002	0.007**
ELA and height	0.013***	−0.006*

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 σ_u and σ_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 (ϕ_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

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

SD of estimated teacher effects–grade 5

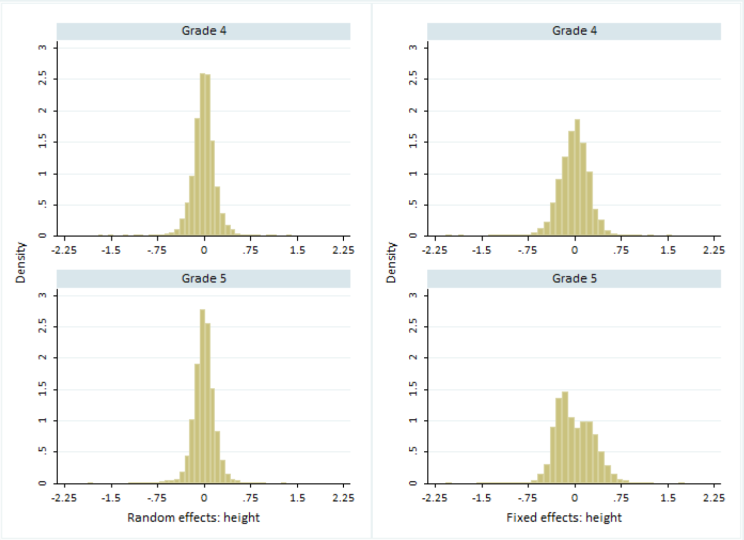
Model:	Height	Math	ELA
A. Baseline models			
RE	0.210	0.253	0.210
FE (adj.)	0.315	0.258	0.240
RE w/school effects	0.157	0.199	0.155
FE w/school effects (adj.)	0.160	0.189	0.145
N of teachers	3,687	4,249	3,978
Mean students per teacher	39.0	42.5	39.5

Teacher effects on height

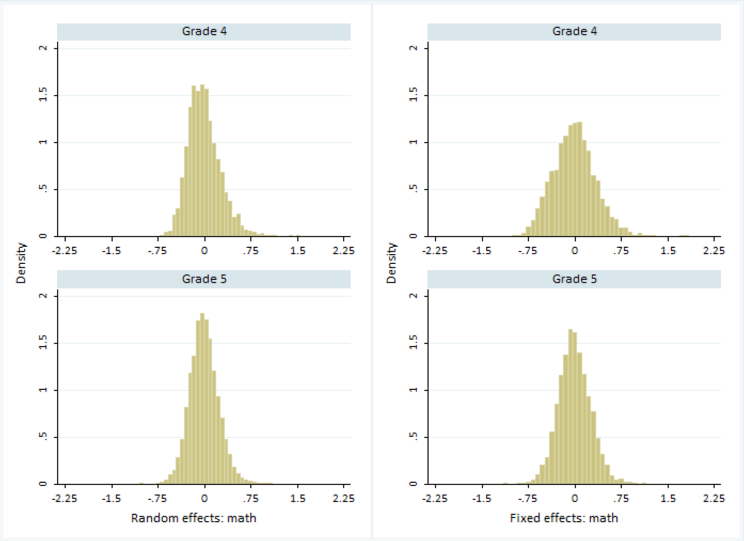
To put in perspective: a 0.22σ increase in height amounts to:

- 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



Distribution of estimated teacher effects—math



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

	Math VAM:			
Grade 4	RE	FE (adj)	RE w/ school effects	FE w/ school effects
Height VAM:				
RE	**-0.019**	-0.014	-0.007	0.008
FE (adj)	-0.030^{+}	**0.199***	-0.022	-0.023
RE w/school effects	0.000	-0.003	**0.002**	0.002
FE w/school effects	-0.002	-0.004	0.001	**0.000**
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

	Math VAM:			
Grade 5	RE	FE (adj)	RE w/ school effects	FE w/ school effects
Height VAM:				
RE	**0.016**	0.015	0.002	0.002
FE (adj)	0.009	**0.090***	0.005	0.005
RE w/school effects	0.001	0.002	**-0.006**	-0.007
FE w/school effects	0.000	0.002	0.005	**0.005**
ELA VAM:				
RE	**0.557***	0.540*	0.438*	0.434*
FE (adj)	0.511*	**0.562***	0.382*	0.378*
RE w/school effects	0.425*	0.406*	**0.514***	0.509*
FE w/school effects	0.424*	0.405*	0.514*	**0.511***

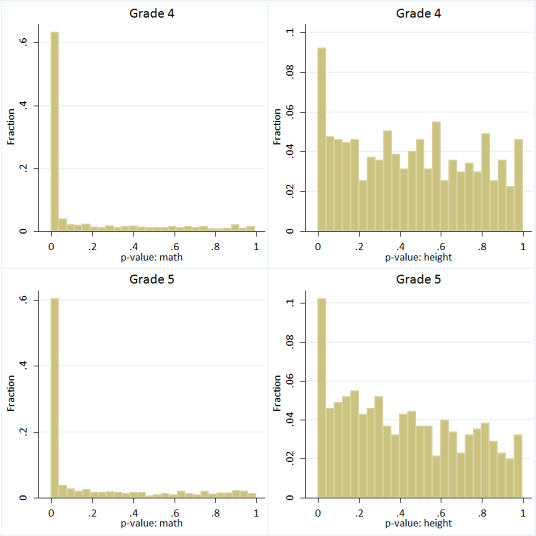
Tests for tracking on lagged student characteristics

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 ($u_{jt} = u_j + v_{jt}$)
- σ_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

	Grade 4	Grade 5	N(4)	N(5)
Height:				
RE	-0.166	-0.167	3,319	3,135
FE (adj)	0.001	-0.094	3,319	3,135
RE w/school effects	-0.004	0.007	3,285	3,100
FE w/school effects (adj)	0.000	0.011	3,285	3,100
Math:				
RE	0.557	0.479	4,001	3,885
FE (adj)	0.587	0.498	4,001	3,885
RE w/school effects	0.463	0.435	3,988	3,868
FE w/school effects (adj)	0.471	0.438	3,988	3,868
ELA:				
RE	0.456	0.408	3,428	3,357
FE (adj)	0.501	0.453	3,428	3,357
RE w/school effects	0.247	0.210	3,410	3,345
FE w/school effects (adj)	0.249	0.214	3,410	3,345

SD of estimated teacher effects–3-level models

	Grade 4			Grade 5		
Model	Height	Math	ELA	Height	Math	ELA
B. 3-level models (KS&R)						
RE	0.000	0.163	0.104	0.000	0.132	0.097
RE w/school effects	0.000	0.107	0.077	0.002	0.087	0.062
C. 3-level models (MLE)						
RE	0.000	0.199	0.159	0.000	0.164	0.121
RE w/school effects	0.000	0.108	0.070	0.000	0.089	0.056

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 ($\widehat{\sigma_u}$) on each iteration.

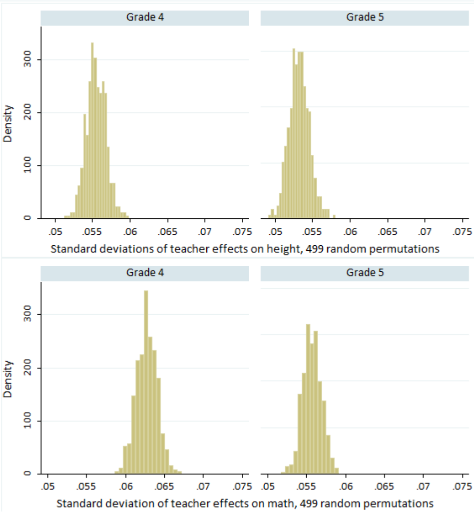
Permutation tests

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 ($\widehat{\sigma_u}$) on each iteration.

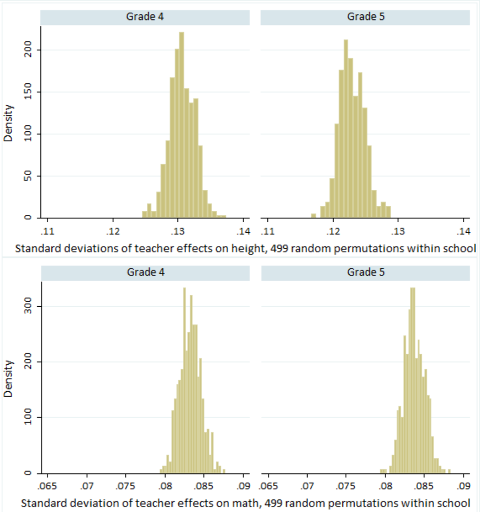
The distribution of these will be informative about the null: what we would expect to see if there were no effects of any kind.

σ_u from permutations—height and math



95th percentiles: 0.058 and 0.056 (height), 0.065 and 0.058 (math)

σ_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.