center on reinventing public education

Teacher Quality and Value-added Measurement

Dan Goldhaber

University of Washington and The Urban Institute

dgoldhab@u.washington.edu

April 28-29, 2009

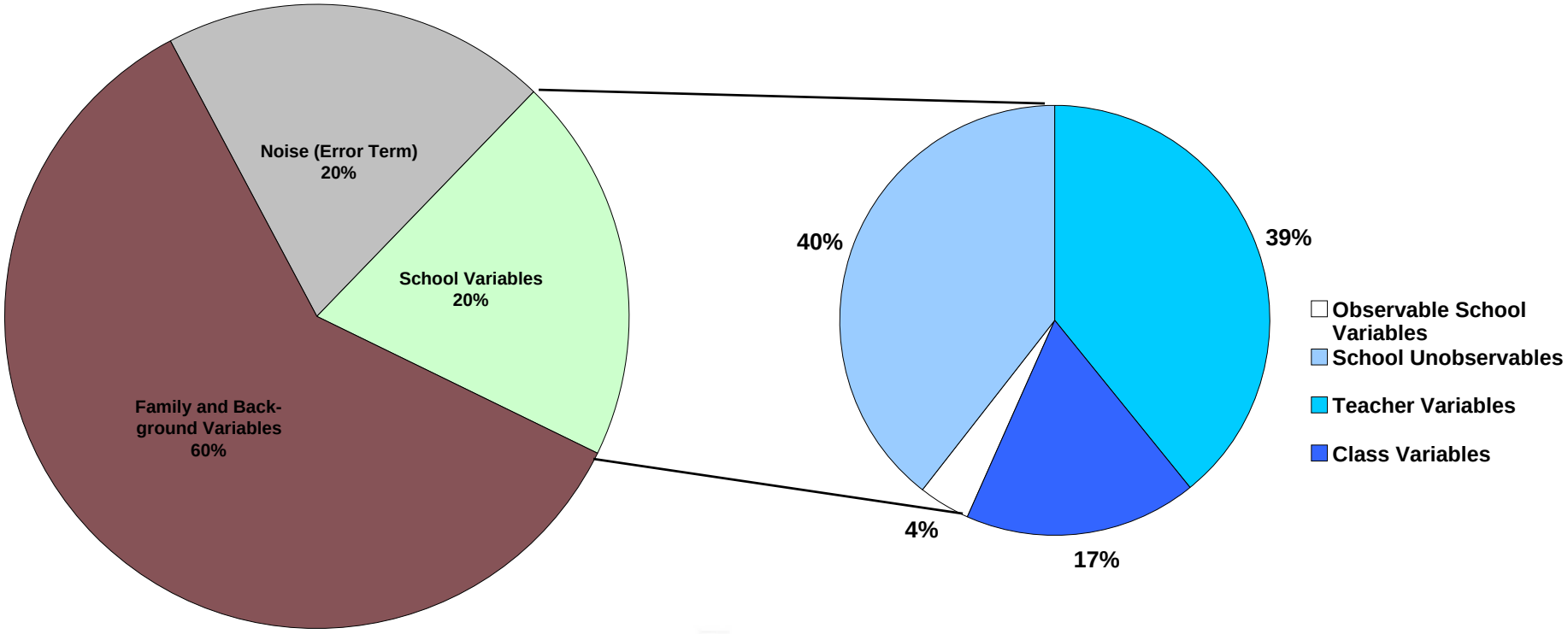
Prepared for the TQ Center and REL Midwest Technical Assistance Workshop:

Evaluating Teacher Effectiveness: The What, How and Why of Educator Evaluation

We Know Teachers Matter!

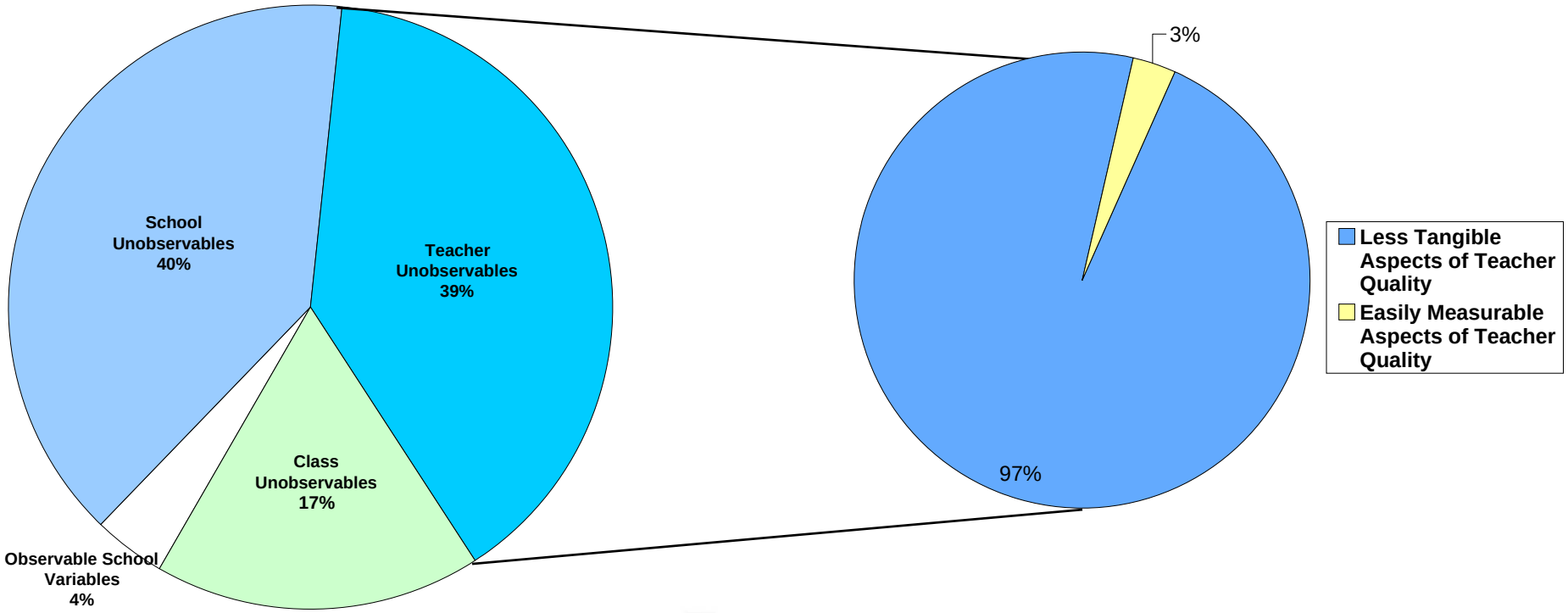
- Controlling for family background factors, teacher quality is the single most important schooling factor explaining student achievement
  - Teacher quality can explain more than one grade-level equivalent in test performance (Hanushek,1992)
  - Impacts of teacher quality can persist for many years (Sanders and Rivers, 1996)
  - Tremendous variation in teacher effectiveness (Bembry et al., 1998; Hanushek, 1992; Sanders and Rivers, 1996)
  - Impact of teacher quality is far larger than any other quantifiable schooling input (Goldhaber, 2002)

Teacher Quality Appears to be Primarily "Unobservable"



Source: Goldhaber et al., 1999

Teacher Quality Appears to be Primarily "Unobservable"



Source: Goldhaber et al., 1999

What Policy Debates Arise From Teacher Quantity Challenge?

- Proper role of state regulation of entry into teaching profession
  - Abel, Fordham, Darling-Hammond, Ballou and Podgursky debates
- Level and structure of teacher salaries
  - Increase teacher salaries, restructure compensation, or do both

Teacher Licensure ("Certification")

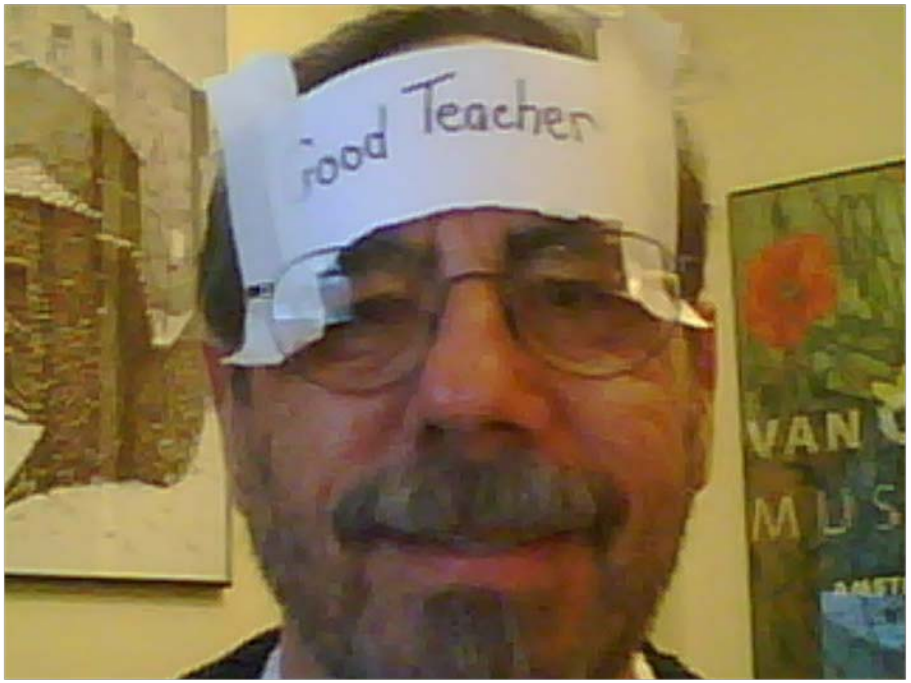
- Licensure system designed to screen out low-quality applicants
  - Completion of approved teacher training program
  - Pre- and post-licensure tests
  - Requirements vary considerably by state
- Debate over licensure system
  - Effectiveness of teachers with standard vs. alternative licensure
  - Increasing standard licensure requirements and closing of "loopholes"
    - Misses the point by ignoring the relevant alternatives for many systems

Licensure Theory

- Protects consumers (ultimately students) from poor choices
  - Localities may make poor or purposeful hiring decisions
    - Bad information or nepotism
- Limits choices of localities and may dissuade talented individuals from considering teaching
  - Localities may have better information than states over who should be hired
  - Limits labor mobility from state to state
- Problem of false negatives and positives

Hypothetical Relationship Between Teacher Licensure-Test Performance & Teacher Quality

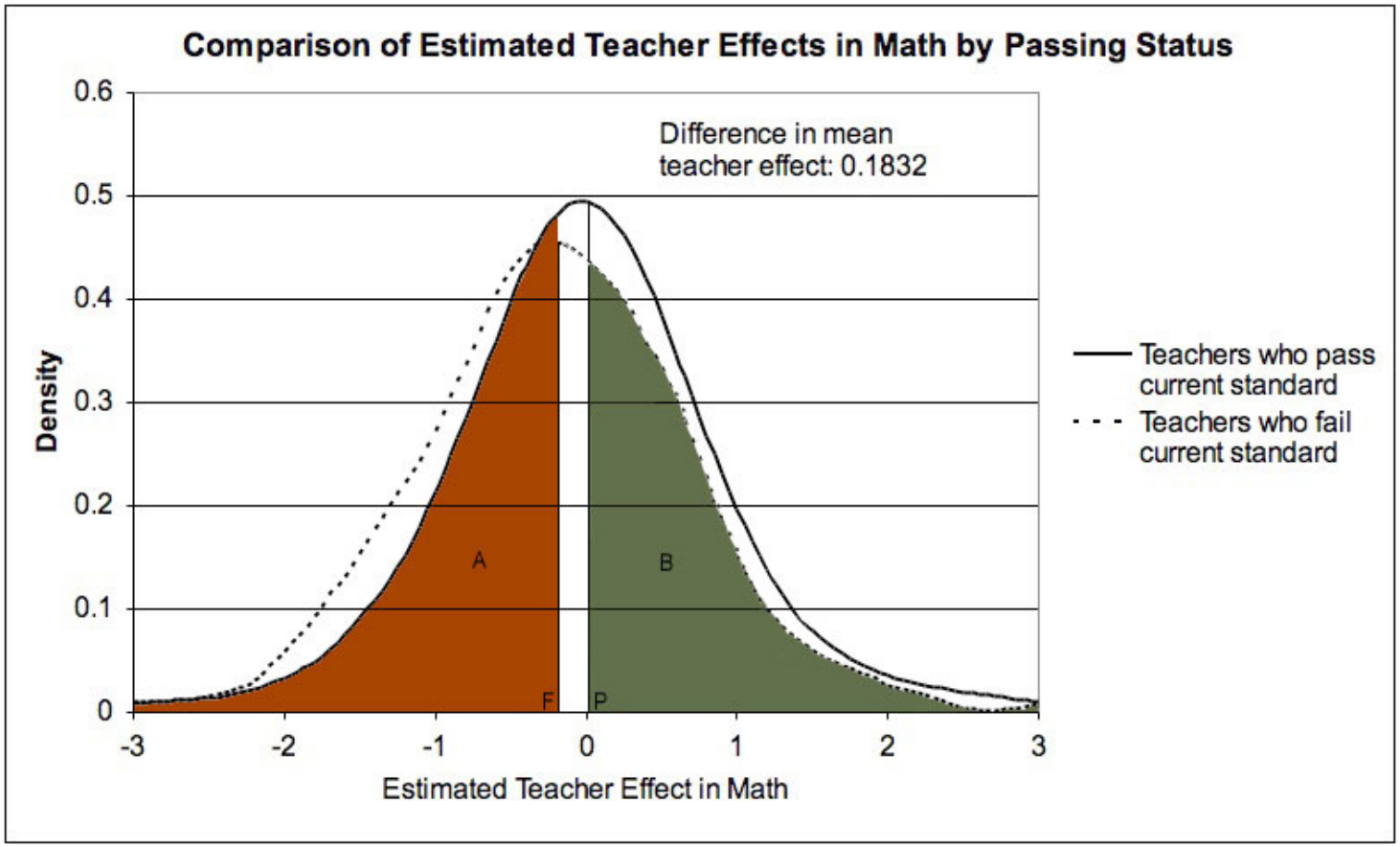
Maybe I'm Wrong!



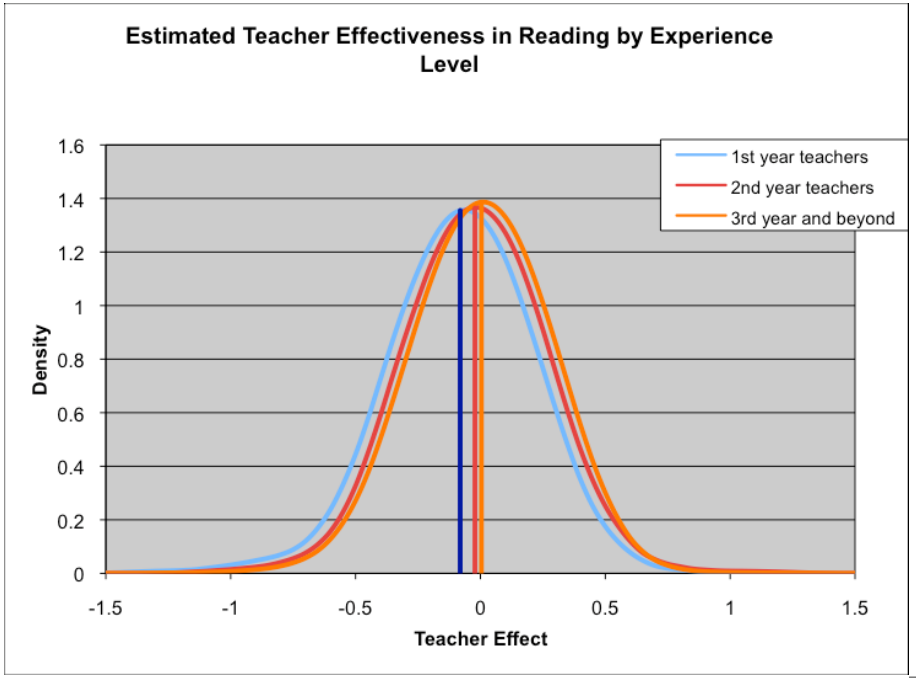
"…We know that teachers are the most important thing,
but teacher quality is not stamped on someone's forehead."

(Dan Goldhaber, *New York Times*, February 22, 2009)

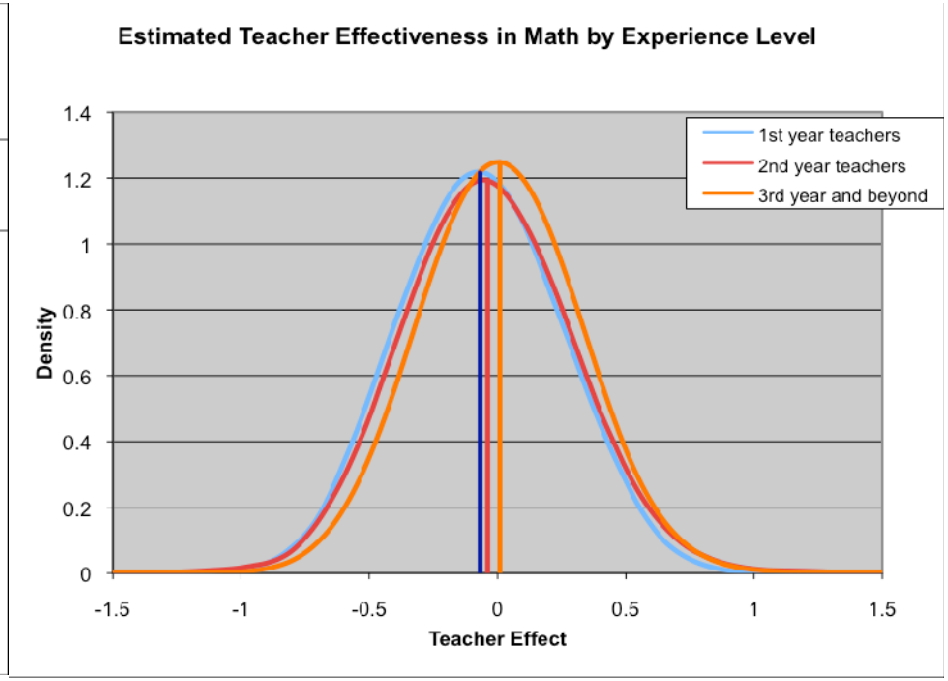
Comparison of Teacher Effects in Math by Passing Status



Experience Levels

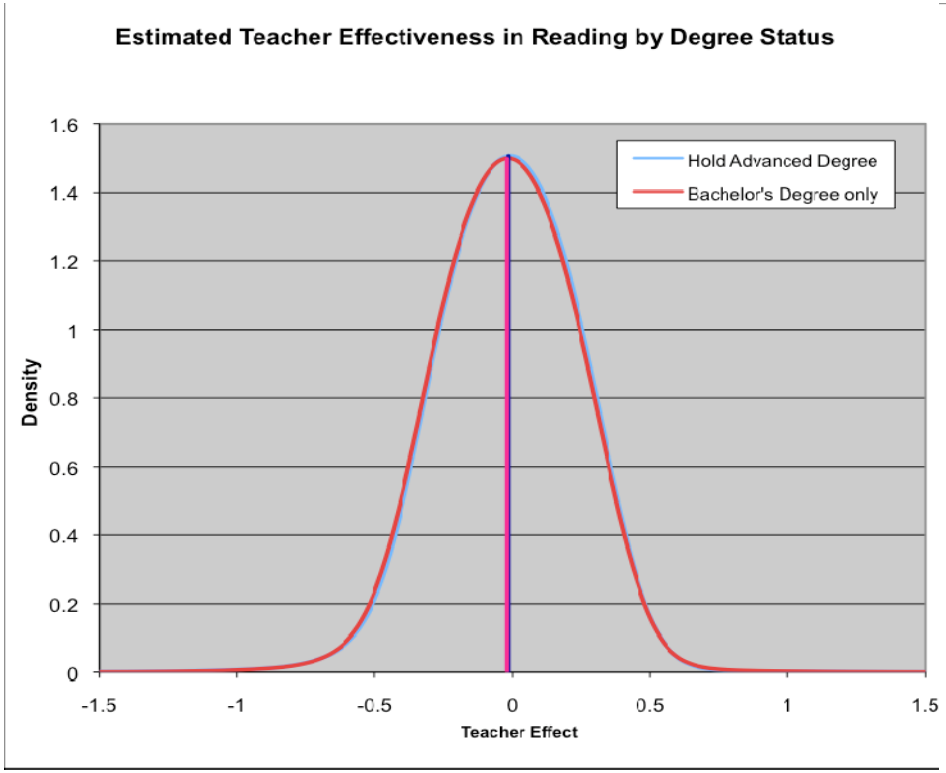


1st year mean-2nd year mean: 0.059** sd
2st year mean-3nd year plus mean: 0.026* sd

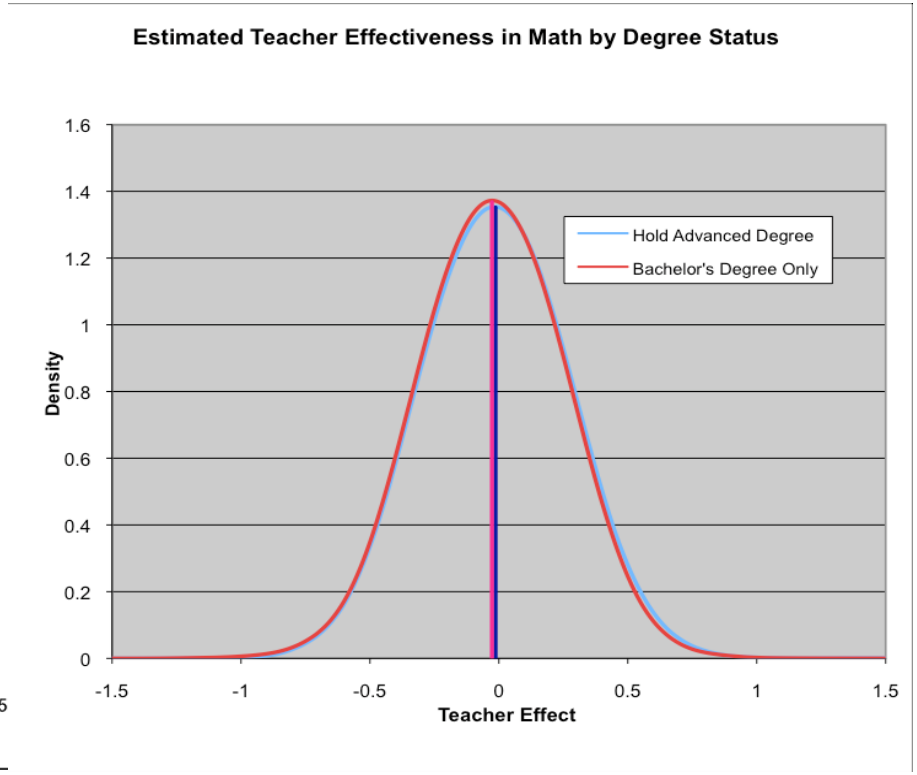


1st year mean-2nd year mean: 0.050* sd
2st year mean-3nd year plus mean: 0.039** sd

Degree Levels

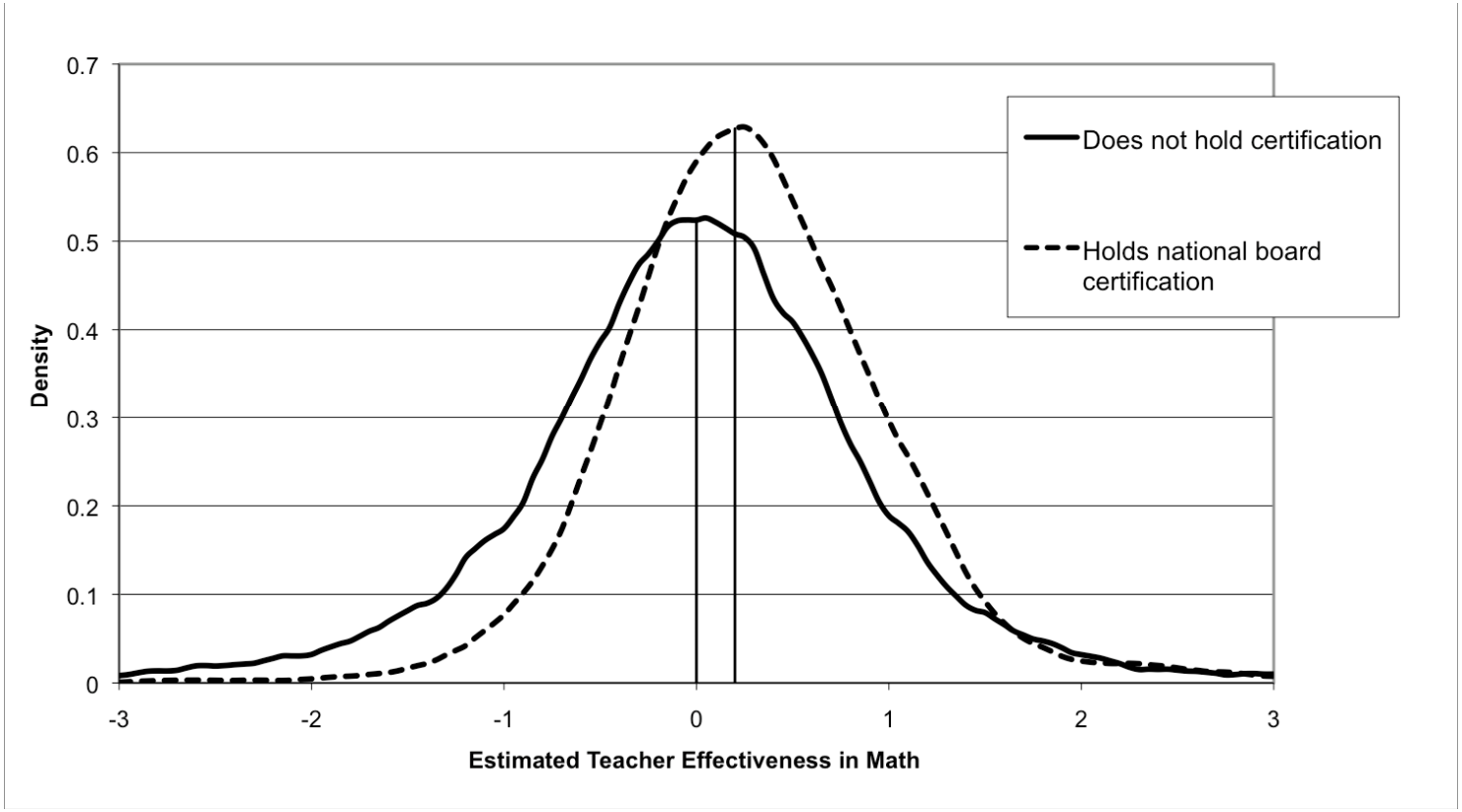


Difference in means: .005 sd



Difference in means: .014 sd

NBPTS Certification Status



Difference in means: 0.19** sd of teacher quality

Arguments for Using VAMs to Assess Teacher Job Performance

- Teachers are the most important *schooling* factor explaining variation in student achievement, but …
  - (Easily quantifiable) teacher characteristics used to determine teachers' employment eligibility and compensation don't strongly predict teacher effectiveness
  - Even when there are statistically significant differences, the differences between the best and worst teachers who hold a particular credential swamp the differences between those with and without the credential

- VAMs may draw different people into teaching, thus helping to address the long-term downward trend in theacademic skills of the U.S. teacher workforce

Using VAMs for Policy Purposes

- Pay, tenure, and teacher "de-selection" reforms
  - Tennessee and Dallas using individual teacher as unit of analysis
  - Pay-for-performance in Florida, Texas, and Minnesota; TIF grantee districts
  - New York City vs. New York State on student test scores
  - De-selection/selective retention ideas associated with researchers (Gorden et al., 2006; Hanushek, forthcoming)
- Underlying tenure/de-selection is the notion that teacher quality is relatively stable characteristic

But… Significant *Potential* Problems with Using VAMs

- Logistical issues (timing of tests; # of tested grades/subjects)
- Perverse incentives/unintended consequences (reclassification of students; too-narrow focus on tested items; discourage collaboration)
- Theoretical/practical issues measuring teacher contributions (cross-subject complements)
- Defining the constructed counterfactual (within or between school/district comparisons)
- **Measurement issues/stability of teacher performance**
  - Signal-to-noise ratio
  - Year-to-year changes in estimated performance
  - Sensitivity of performance ranking to changes in sample, subject, or teaching context

Thoughts on VAMs in Practice

- For policy purposes we probably don't care about precise estimates of teacher effects
  - We care about where in the effectiveness distribution teachers fall
  - VAM estimates can be wrong, but not so wrong that they radically change the estimated teacher-effectiveness distribution
  - We don't know much about how or whether VAM errors influence where teachers fall in the distribution
- Are we holding VAMs to a higher standard?
  - Estimates of productivity may be as imprecise and vary as much in the private sector

Focus of this Work

Assess the stability of (value-added) teacher job performance estimates over time, including a focus on pre- and post-tenure

North Carolina Data

- Administrative records for all NC teachers and students for grades 3-8 from 1995-96 to 2005-06
  - Fifth-grade performance for students with full history of test scores & in classes with 10-29 students
- Track teachers for whom we observe for at least two years pre-tenure and one year post-tenure
  - 281 unique teachers in this select sample

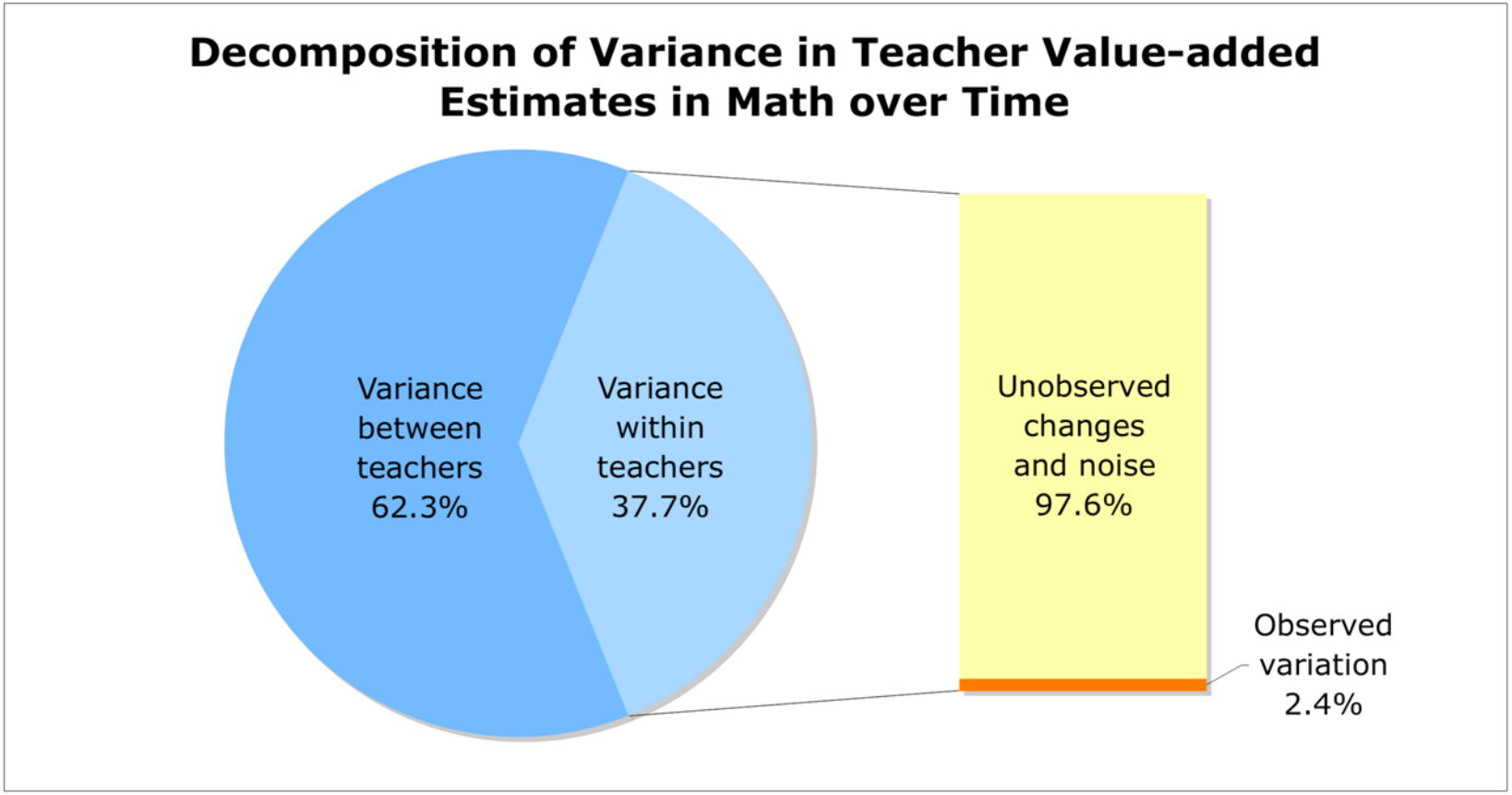
Analytic Approach

- $A_{i,j,t,s,g=5} = \alpha \mathbf{A}_{\mathbf{i(history)}} + X_{i,t,g=5}\gamma + \tau_{j,t,g=5} + \varepsilon_{i,j,t,s,g=5}$ where $\mathbf{A}_{\mathbf{i(history)}} = \left[A_{i,R,g=4} \middle| A_{i,M,g=4} \middle| A_{i,R,g=3} \middle| A_{i,M,g=3}\right]$
- Specification is consistent with the unbiased estimates from Kane and Staiger (2008) and the bias-minimizing specification in Rothstein (2008)

Teacher Effects Estimates

- One standard deviation increase in TQ is estimated to increase student achievement by .2 standard deviations (which is approximately 30-40% of the average yearly gain in achievement, so equivalent to about 3 months of learning)
- Variation between teachers explains 52% of overall variance in teacher effects in reading and 63% in math
- Decomposition of teacher effects shows time-varying teacher characteristics explain only a trivial proportion of the variation in the teacher effect estimates
- Average correlation of teacher job performance is 0.32 in reading and 0.54 in math
  - Estimates of stability of job performance are not terribly different from private sector estimates

Components of Estimated Year-By-Year Teacher Effects



Transition Matrices on Adjacent-Year Quintile Rankings

Panel A. Reading Performance

Quantile in Year *t*	% of Total Teachers in Quantile in Year *t*+1					
	1	2	3	4	5	Total
1	5.82	4.39	3.67	2.71	1.82	3,197
2	4.20	4.53	4.10	3.90	2.87	3,440
3	3.37	4.07	4.65	4.35	3.91	3,500
4	2.72	3.44	4.28	4.91	4.91	3,603
5	1.72	2.95	3.72	4.98	8.02	3,717
Total	3,138	3,358	3,570	3,619	3,772	17,457

Panel B. Math Performance

Quantile in Year *t*	% of Total Teachers in Quantile in Year *t*+1					
	1	2	3	4	5	Total
1	7.63	4.73	3.28	1.91	0.77	3,213
2	4.78	5.22	4.55	3.47	1.67	3,421
3	3.08	4.34	4.91	4.61	3.11	3,551
4	1.86	3.34	4.79	5.45	5.21	3,538
5	0.63	1.60	2.92	5.30	10.84	3,734
Total	3,112	3,383	3,563	3,640	3,759	17,457

Pre- and Post-Tenure Job Performance Rankings: Reading

Panel A. Using first two years of performance to predict post-tenure performance

	Post-tenure Quintile Rank					
Pre-tenure Quintile Rank	Bottom Quintile	Second Quintile	Third Quintile	Fourth Quintile	Top Quintile	Total Teachers
Bottom Quintile	32%	23%	19%	16%	11%	57
Second Quintile	27%	14%	27%	18%	14%	56
Third Quintile	21%	23%	30%	18%	7%	56
Fourth Quintile	16%	27%	18%	18%	21%	56
Top Quintile	5%	13%	5%	30%	46%	56
Total Teachers	57	56	56	56	56	281

Panel B. Using first three years of performance to predict post-tenure performance

	Post-tenure Quintile Rank					
Pre-tenure Quintile Rank	Bottom Quintile	Second Quintile	Third Quintile	Fourth Quintile	Top Quintile	Total Teachers
Bottom Quintile	26%	30%	18%	14%	12%	50
Second Quintile	28%	14%	38%	12%	8%	50
Third Quintile	26%	24%	16%	22%	12%	50
Fourth Quintile	12%	18%	22%	24%	24%	50
Top Quintile	8%	14%	6%	28%	44%	50
Total Teachers	50	50	50	50	50	250

Pre- and Post-Tenure Job Performance Rankings: Math

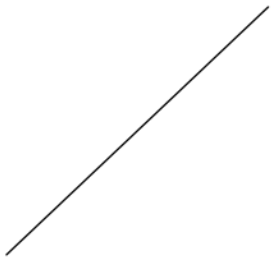
Panel A. Using first two years of performance to predict post-tenure performance

	Post-tenure Quintile Rank					
Pre-tenure Quintile Rank	Bottom Quintile	Second Quintile	Third Quintile	Fourth Quintile	Top Quintile	Total Teachers
Bottom Quintile	44%	25%	14%	16%	2%	57
Second Quintile	25%	30%	25%	13%	7%	56
Third Quintile	14%	14%	30%	18%	23%	56
Fourth Quintile	14%	18%	18%	23%	27%	56
Top Quintile	4%	13%	13%	30%	41%	56
Total Teachers	57	56	56	56	56	281

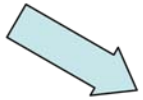
Panel B. Using first three years of performance to predict post-tenure performance

	Post-tenure Quintile Rank					
Pre-tenure Quintile Rank	Bottom Quintile	Second Quintile	Third Quintile	Fourth Quintile	Top Quintile	Total Teachers
Bottom Quintile	42%	26%	18%	10%	4%	50
Second Quintile	36%	28%	20%	12%	4%	50
Third Quintile	16%	24%	26%	18%	16%	50
Fourth Quintile	4%	14%	20%	28%	34%	50
Top Quintile	2%	8%	16%	32%	42%	50
Total Teachers	50	50	50	50	50	250

De-selecting Poor Performers in Either Subject



De-selecting Poor Performers in Both Subjects



Tradeoffs

- Multiple years of job performance data certainly improves reliability of estimates
  - More information & ability to use more sophisticated statistical approaches
    - But, no VAM information on first-year teachers & potential dampening of performance incentives
- Comparisons within and between schools
    - May be few good within district comparisons (in small districts) but allows districts to implement policies (sample issue)
    - Within and between school comparisons conflate school and teacher effects but effective teacher in one school might have been ineffective in another (statistical approach issue)
  - Decisions about comparisons have potentially important policy implications for level of policy implementation
    - States could assist by estimating VAMs, but leaving it up to localities to decide how to use the estimates

In the Eye of the Beholder

- Year-to-year job performance estimates are modest (0.3 in reading and 0.5 in math); pre- and post-tenure estimates are somewhat higher (0.4 in reading and 0.6 in math)
  - We can't know whether these fluctuations represent true changes in job performance
- Inter-temporal estimates are not out of line with those found in other sectors of the economy that use them for policy purposes; and pre-tenure estimates clearly do predict estimated post-tenure performance
- More holistic assessment (complementing VAMs) would be nice, but…
  - Structural impediments to serious evaluation
  - Mistrust of subjective judgments
- How did we get here?
  - Poor evaluation/little use of evaluation today
  - Policymakers hope: VAMs are objective evaluation tool, which allows schools to do what they did not do when left to their own devices
- More research needed on using VAM to identify individual teacher effectiveness
  - Perfect can be the enemy of the good; we cannot learn all of what we need to know outside of actual policy variation

For More Detail…

- www.crpe.org
- www.caldercenter.org
- Goldhaber Dan and Hansen, Michael. “Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance.” CRPE Working Paper #2008-5. (November 2008).
- Goldhaber Dan and Hansen, Michael. “Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions.” CRPE Research Brief (November 2008).
- Sass, Tim R. “The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy.” Presented at the Second Annual CALDER Conference (November 2008).

VAM Discussion Questions

1. Are student tests important measures of learning?
2. How should we evaluate teachers in non-tested subjects/grades?
3. What are the ways of mitigating perverse incentives/unintended consequences
4. What are the right VAM teacher comparisons?
5. How much teacher-student information is enough to make judgments about teachers?