The Promise and Potential Pitfalls of Value-Added Assessment

Dan Goldhaber
Center on Reinventing Public Education
University of Washington

Just in Case…

- There are potentially a lot of problems with using VAM for policy purposes
  - Inadequate teacher quality measures and inappropriate uses of measures may create perverse incentives and certainly political strife
- There's a fair amount that we don't know about using VAM
- But, if I had to make a call
  - Just as "democracy is the worst form of government except all the others that have been tried"… VAM isn't perfect, but it looks good relative to alternatives (at least with the current institutional structure/culture of public schools)

A Simple Observation

- Opposition to VAM arises not from intrinsic opposition to statistical approach, but in how it might be used
- Thus, this presentation is focused on VAM in the context of it's potential uses
  - Directing professional development
  - Factor in pay
  - Determining employment

Do We Care About Test Scores?

- The problem: all tests sample imprecisely over various domains
- But, there is a large literature showing that various tests predict
  - College-going behavior, employment probability, earnings, and a host of other non-financial measures of well-being
  - National competitiveness (Hanushek et al., 2008)

→ Some question of causality, but I would argue the evidence is pretty definitive that tests are important measures

Gateway vs. Workforce Policies

Empirical arguments for teacher workforce policies:

1. (Easily quantifiable) teacher characteristics used to determine teachers' employment eligibility and compensation don't strongly predict teacher effectiveness
2. Teachers are more different than alike: the differences between the best and worst teachers who hold a particular credential swamp the differences between those with and without the credential

But… Significant *Potential* Problems with Using VAM

- Logistical issues
- Perverse incentives/unintended consequences
- Test measurement issues
- Theoretical/practical issues measuring teacher contributions
- Defining the constructed counterfactual

Logistical Issues

- Timing of tests
  - Summer fall-back
  - Administration mid-year
- Student absences
- Data constraints
  - In most places data systems are not now set up for high-stakes VAM purposes
    - To get more credible estimates of teacher contribution we need multiple years of student background information linked to teachers
    - To use VAM effectively, we probably need VAM estimates sooner

Perverse Incentives/Unintended Consequences

- Forgo non-tested outcomes (history, music, tolerance, democracy, etc.)
  - Less than 20% of K-12 instruction is in tested areas
- Narrowing of instruction to *unimportant* test skills not associated with comprehension
  - We often see large gains on state assessments when a new test is introduced (not confirmed by NAEP)
- Discourage collaboration amongst teachers
- Corruption
  - Hiding of low-growth students
  - Cheating

Test Measurement Issues

- Vertical alignment of tests & floor and ceiling effects
- Tests are noisy measures of student knowledge
  - Precision of estimate of teacher value added will depend on number of students taught (class size and/or # years of data in VAM)
    - Kane & Staiger (2001) find probability of falling into tails of the distribution is inversely related to school size (small schools more likely to be rewarded or sanctioned because of statistical noise)
    - Estimates (e.g. Ballou, 2005) suggest that only 2.5 (reading) and 17 (math) percent of 1-year teacher effect estimates are statistically significant, but 3-year estimates roughly triple the percentage of statistically significant effects
  - Imprecision means that "cut-point policies" will *always* result in errors

Measuring Teachers

- We don’t really know how to handle:
  - Teachers who are effective in one area and not another
  - Multiple teachers, complementary subjects and apportioning credit
  - Out of grade teacher contributions

Defining the Counterfactual

- We never directly observe the counterfactual of interest
  - VAM should account for:
    1. Observed differences in student background/skills
    2. Observed differences in school quality and peer effects
    3. Unobservables related to the nonrandom distribution of students and teachers
- Failure to account for any of the above three issues can lead to biased VAM estimates

What We Do Know About VAMs?

- Relatively few teacher effects are statistically different from mean effect
- Year-to-year correlation of teacher effects is in the range of .2 to .3
- VAM models accurately estimate teacher performance in the absence of detailed student background information, but *only if* there is significant mixing of students across teachers (McCaffrey et al., 2004)
- VAMs fail various falsification tests (Rothstein, 2008)
- Experimental VAM estimates look similar to nonexperimental estimates with particular specifications (Kane and Staiger, 2008)

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

Arguments for Using VAM For Pay Purposes

- VAM may draw different people into teaching addressing the long-term downward trend in the academic skills of the U.S. teacher workforce
- If the desire is to reward teachers who produce high value-added then employing a credentials-based strategy will lead to significant errors (experience, degrees, NBCTs)
- Few examples of long-standing programs, but recent empirical work shows that pay for performance increases student achievement (e.g. Figlio and Kenny, 2006; Lavy, 2002, 2004)

Current Salary Structure Isn't Working Well

- To bring most academically proficient individuals into teaching

  On average, teachers:

  – Graduate from less-selective undergraduate institutions, have lower standardized test scores (Ballou, 1996; Goldhaber and Liu, 2003; Hanushek and Pace, 1995), and require more remediation in college (U.S. Department of Education, 1996)

- "College graduates with high test scores are less likely to take [teaching] jobs, employed teachers are less likely to stay, and former teachers with high test scores are less likely to return" (Murnane, et al., 1991)

Thoughts on VAM 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 VAM to a higher standard?
  - Estimates of productivity may be as imprecise and vary as much in the private sector

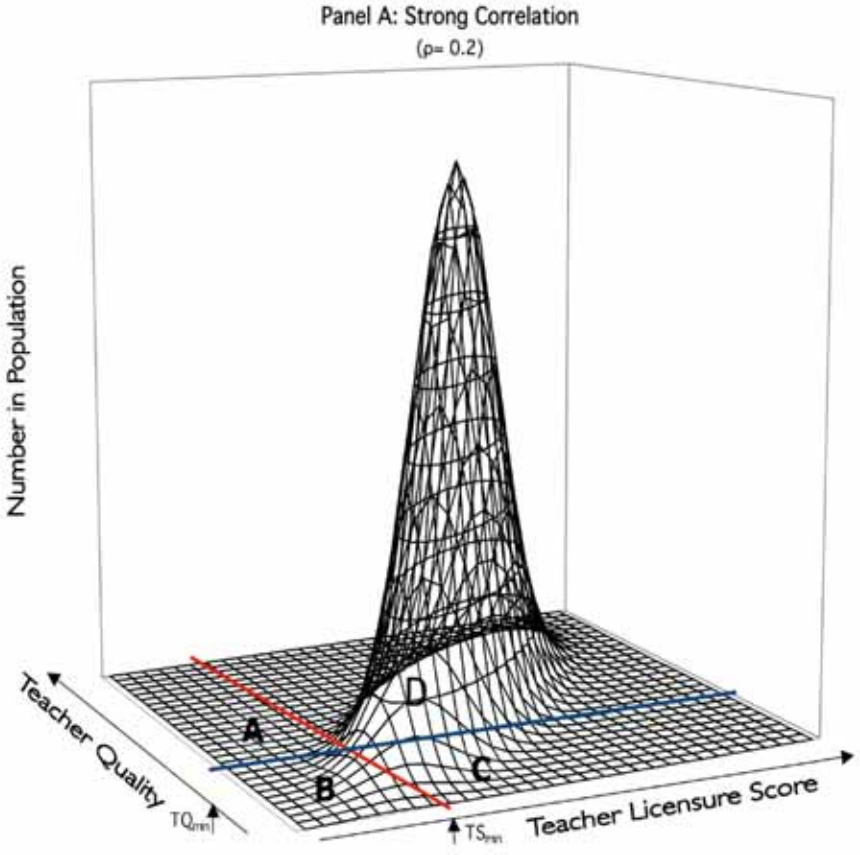
Why VAM?
Devil You Don't Know Is Preferable

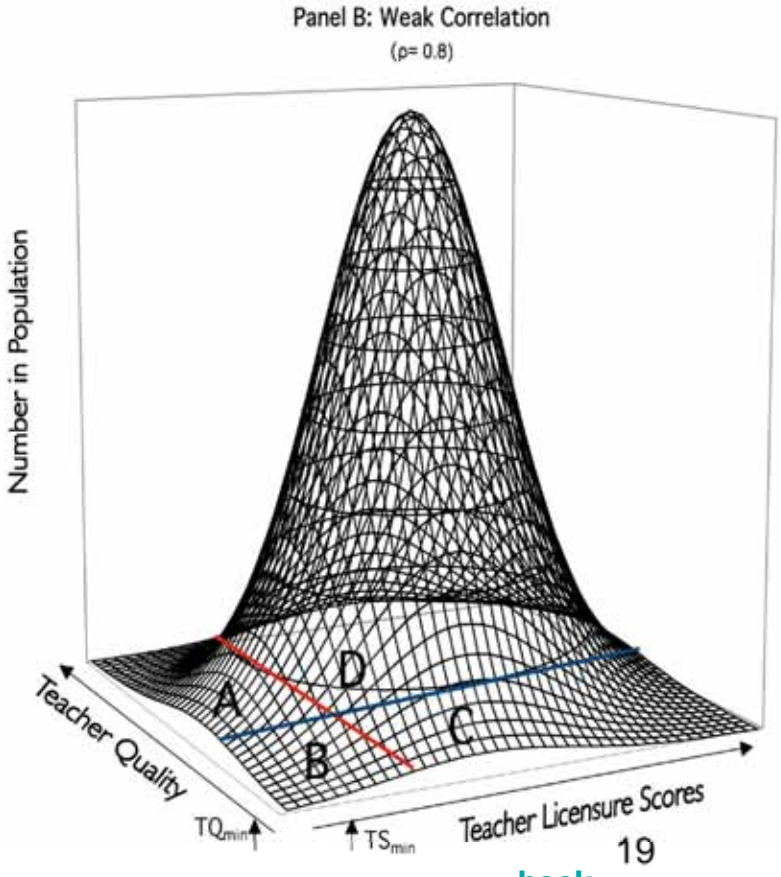
- We need to try different approaches
  - Substantial evidence that we currently make bad investments
    - MA pay premium
    - Professional development
  - Stubborn problems with student achievement
    - Only spotty and modest improvement over time, despite significant increases in spending on K-12 schools
    - Large achievement gaps between various groups still exist
    - U.S. students substantially lagging in international comparisons

Alternative is Far from a Utopia

- More holistic assessment (complementing VAM) would be nice, but…
  - Structural impediments to serious evaluation
  - Mistrust of subjective judgments
- How did we get here?
  - Accountability and standards movement: public dissatisfaction with schools (students graduating from high school without basic literacy or numeracy skills)
  - Policymakers hope: VAM is objective evaluation tool, which allows schools to do what they did not do left to their own devices
- I would prefer using VAM, even with potential problems, if the alternative is no serious evaluation (with consequences) of teachers

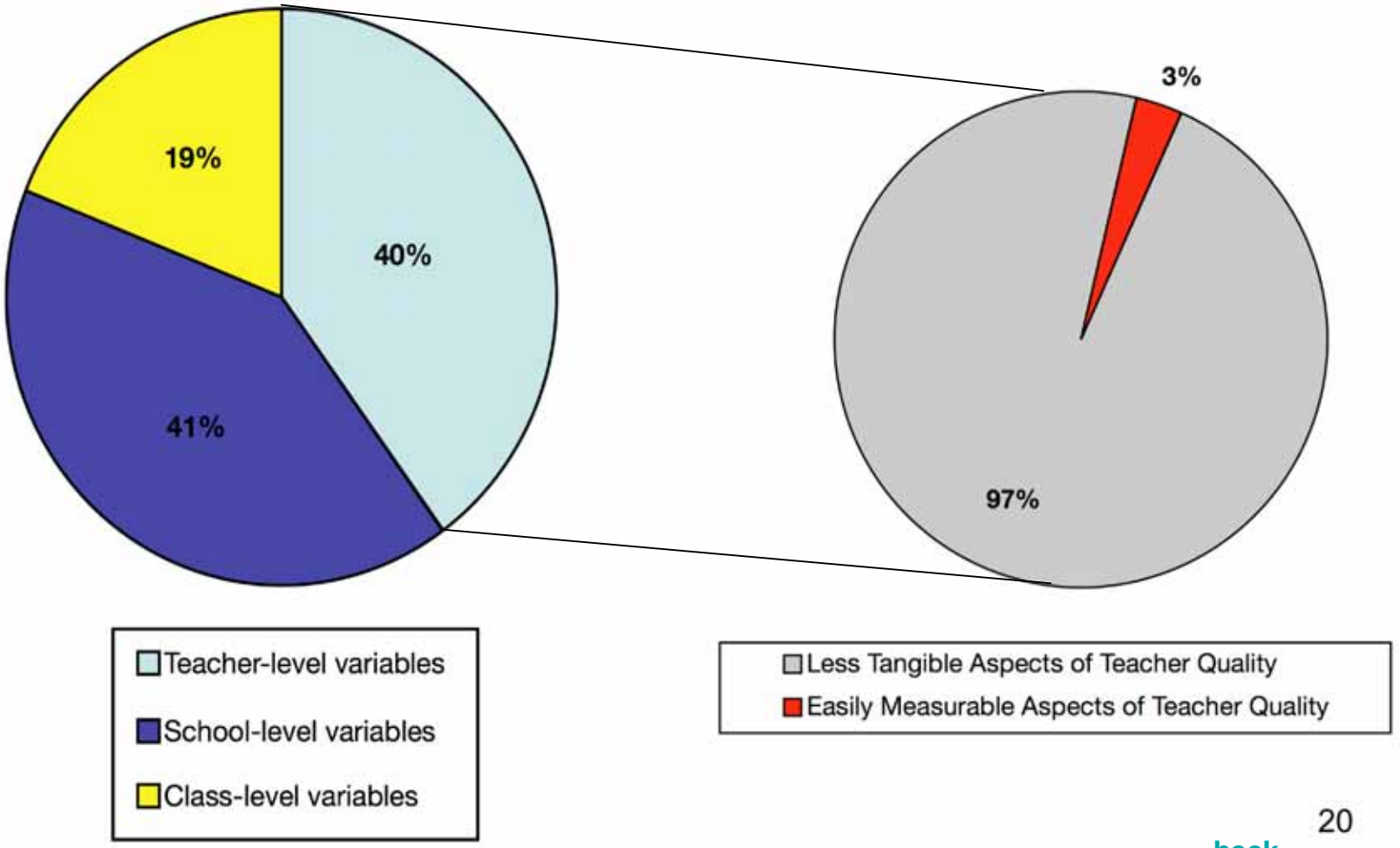
Hypothetical Relationship Between Teacher Licensure Test Performance and Teacher Quality





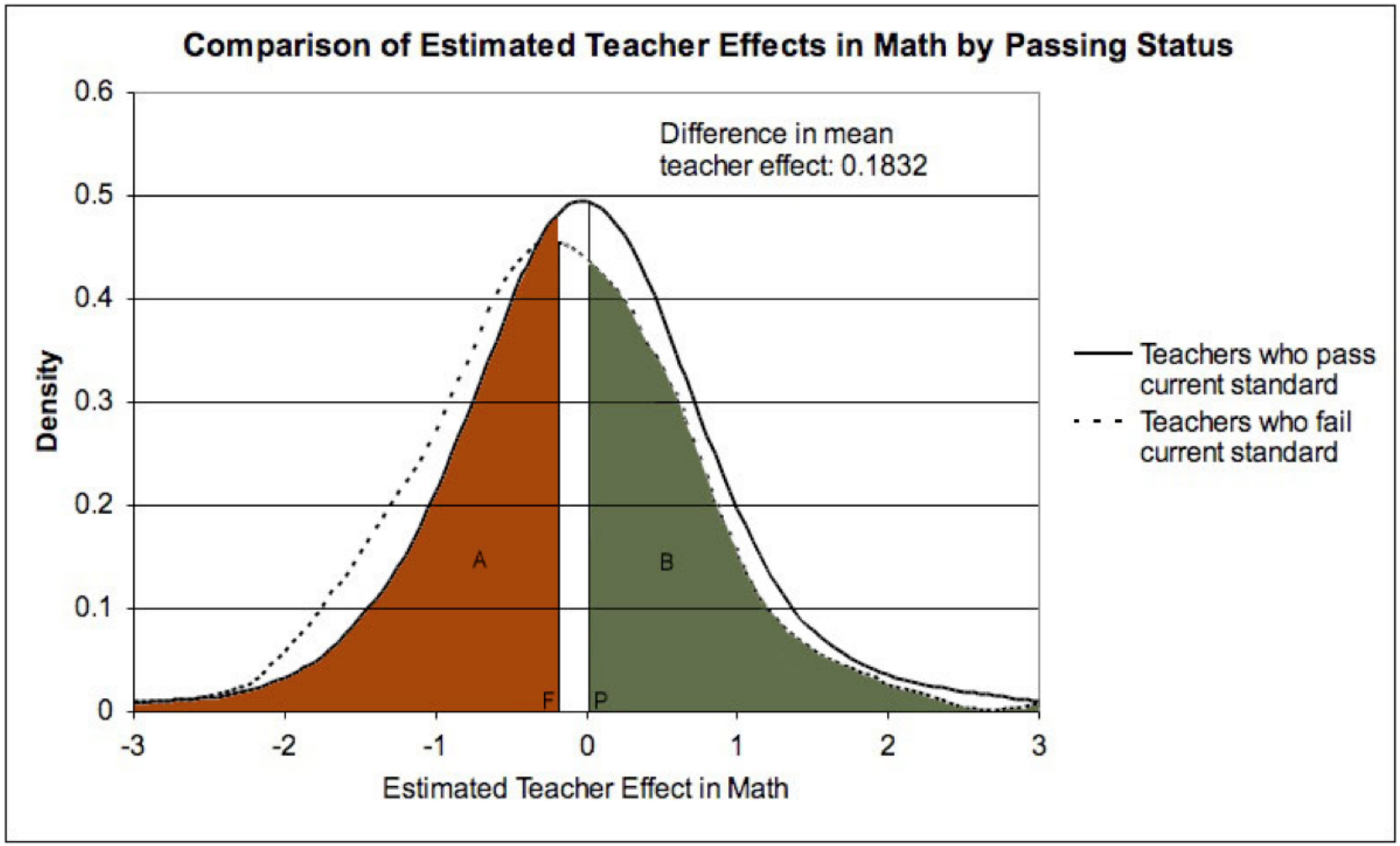
back

Teacher Quality Appears to be Primarily "Unobservable"



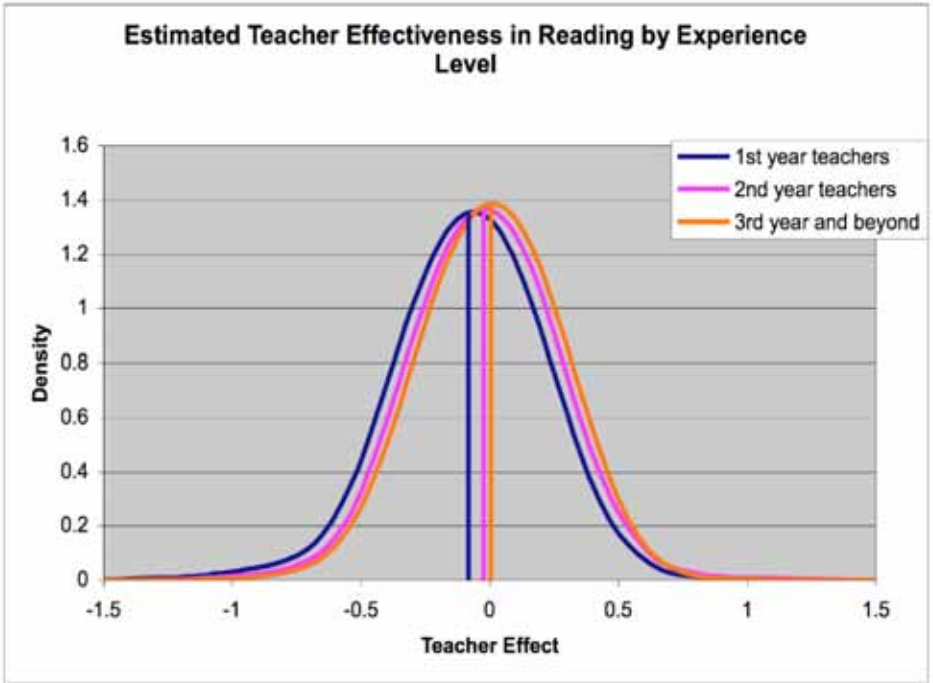
back

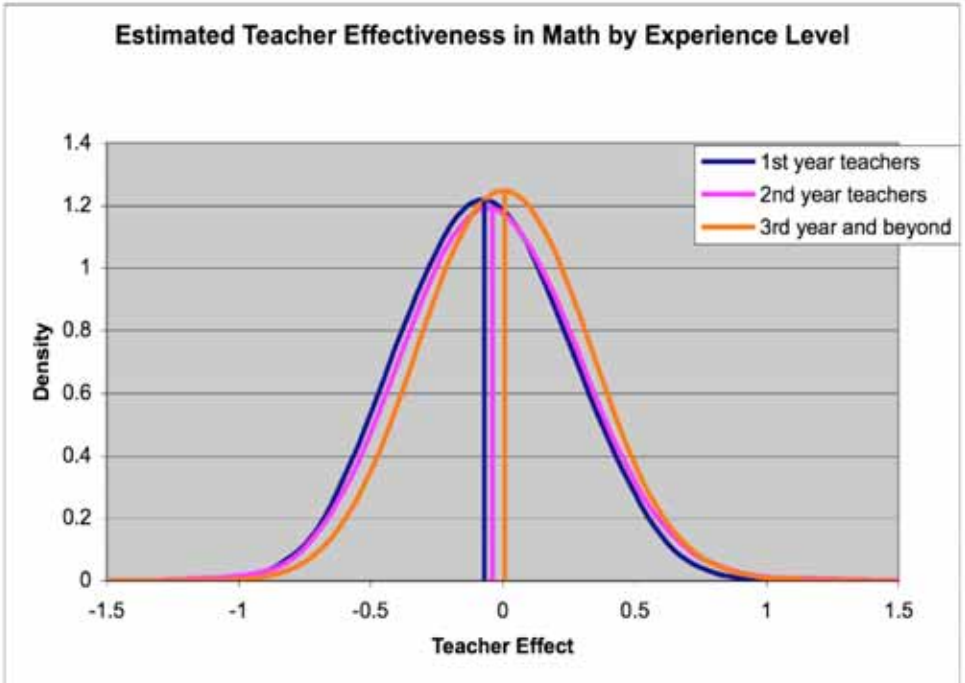
Comparison of Teacher Effects in Math by Passing Status



back

Experience Levels



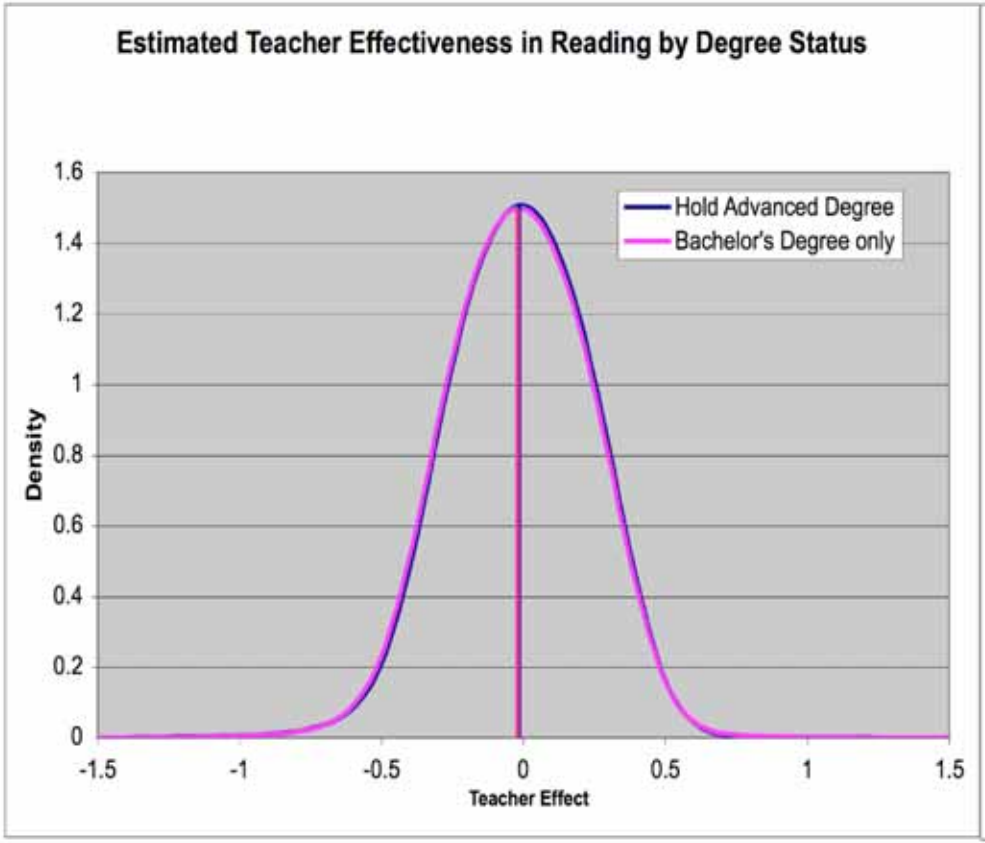


1st year mean-2nd year mean: 0.059** sd
2st year mean-3nd year plus mean: 0.026* sd

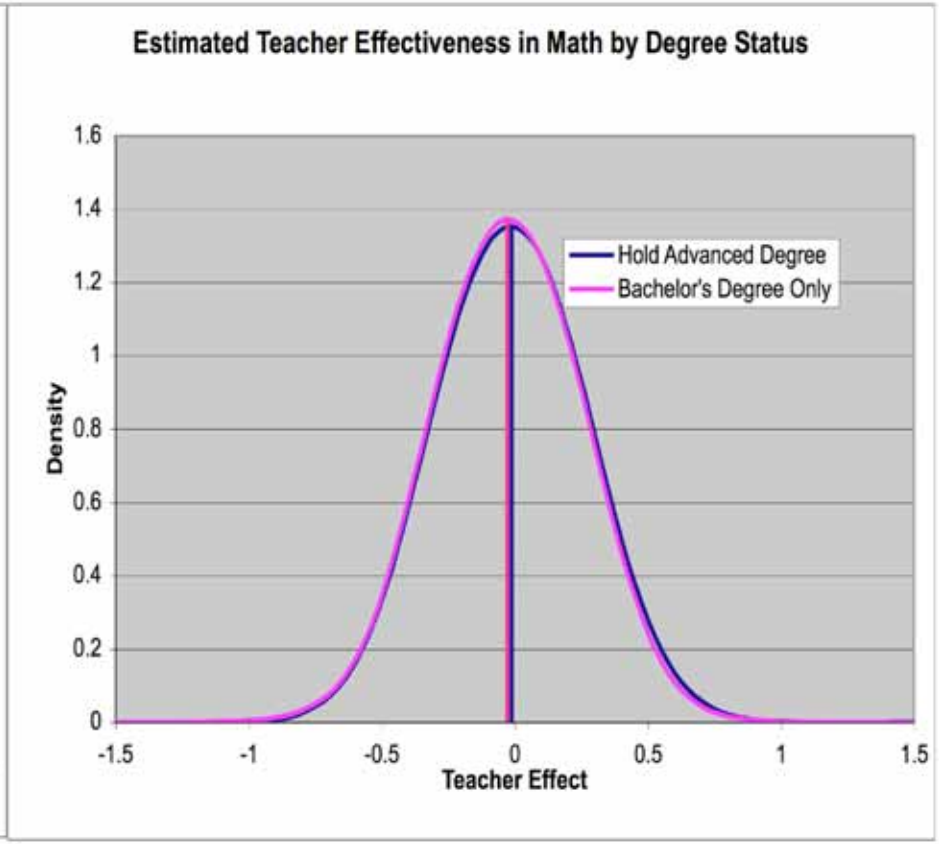
1st year mean-2nd year mean: 0.050* sd
2st year mean-3nd year plus mean: 0.039** sd

back degrees

Degree Levels



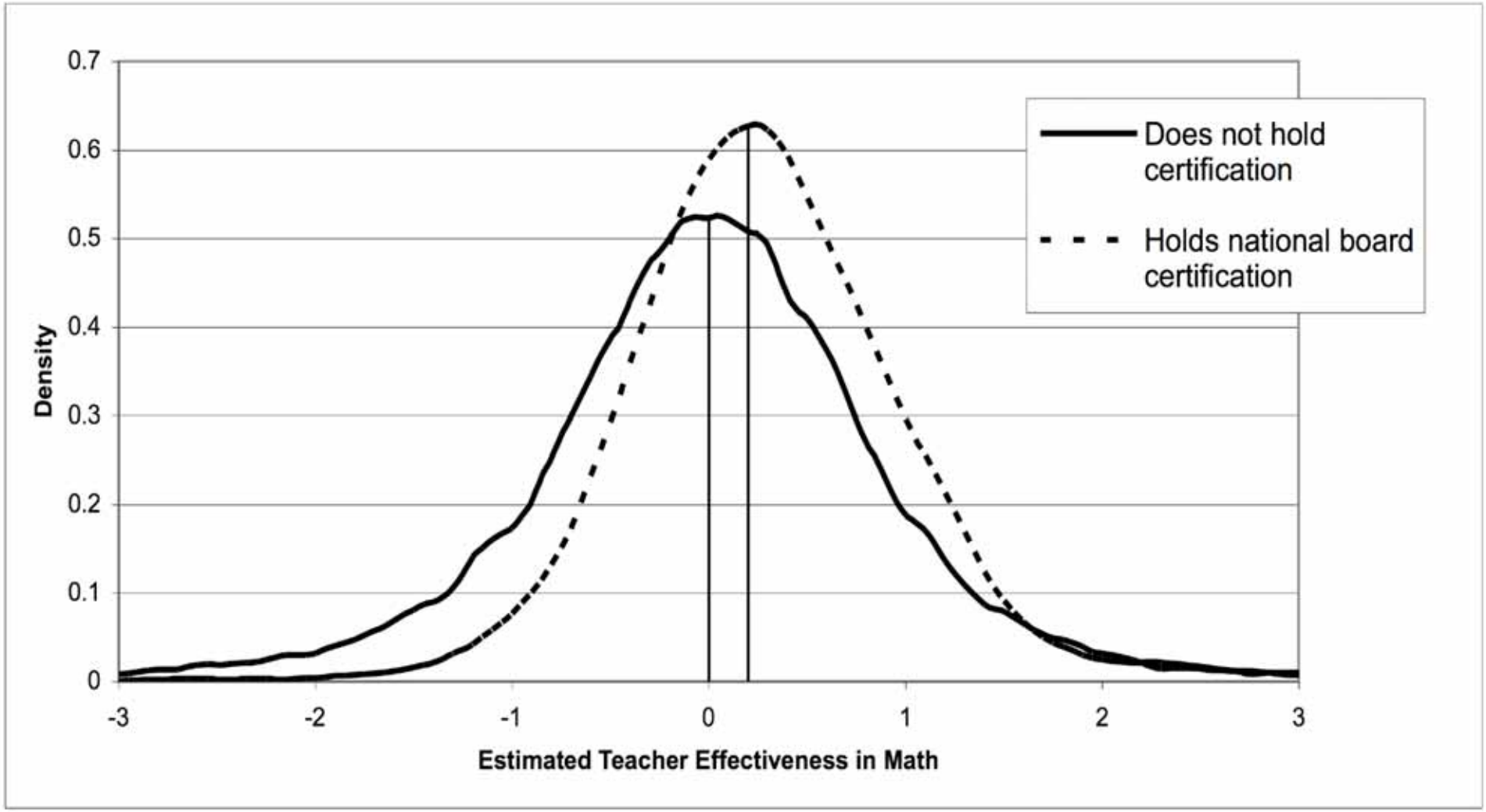
Difference in means: .005 sd



Difference in means: .014 sd

back experience

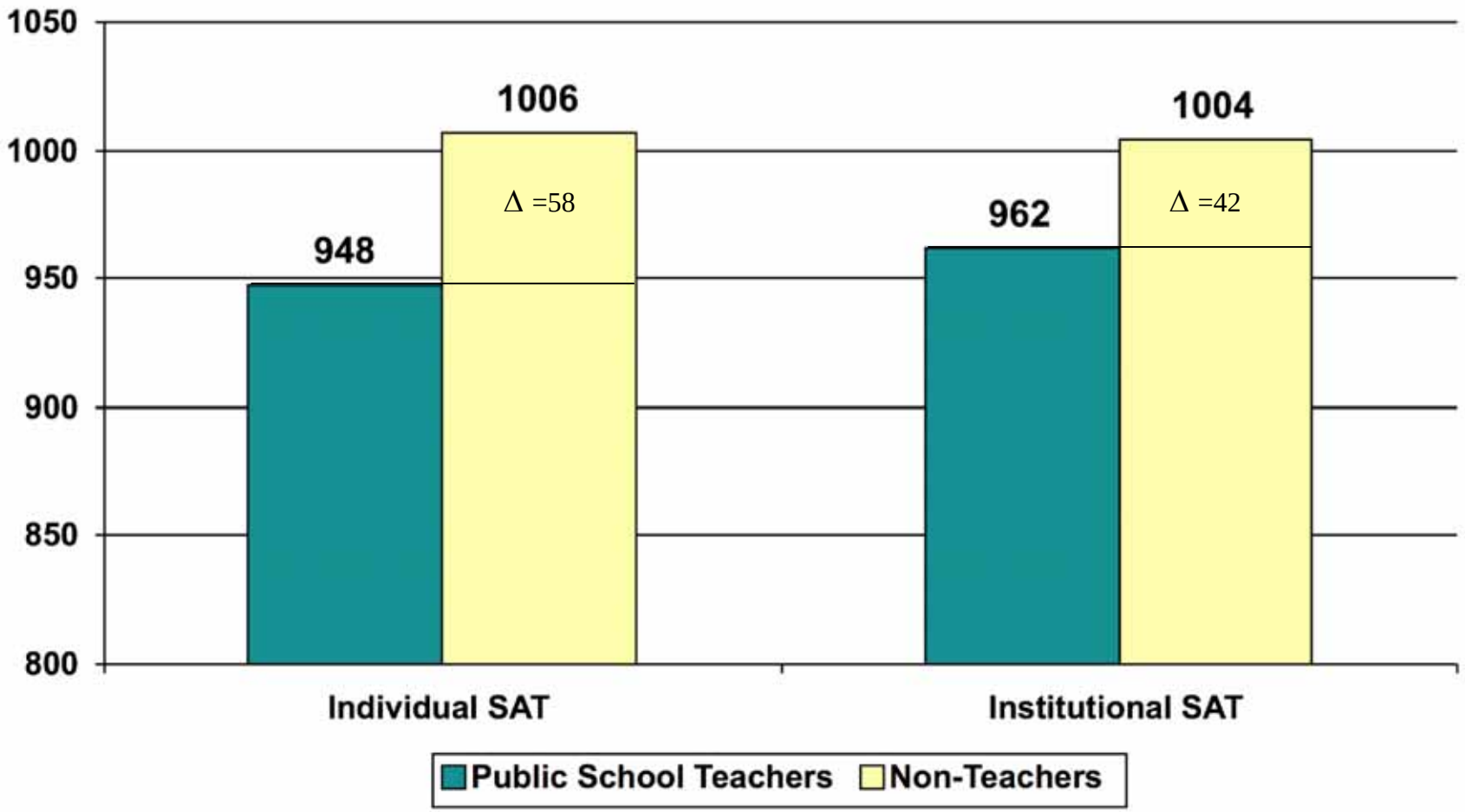
NBPTS Certification Status



Difference in means: 0.19** sd of teacher quality

back degrees experience

Individual & Institutional SAT Scores



Source: *Baccalaureate and Beyond*

back

What Would It Cost To Raise Teacher Salaries To That Of Other Professionals?

	1999 Average Annual Salary	% Increase In Teacher Salary	Total Necessary Spending On Education	Needed Spending Per Pupil
Teacher	$48,689	--	$355 billion	$6,508
Family Physician	$133,900	175%	$597 billion	$12,734
Full Professor	$78,830	62%	$441 billion	$9,412
Attorney	$69,104	42%	$413 billion	$8,826
Engineer	$68,294	40%	$411 billion	$8,777

Source: *AFT Salary Survey 2000*

Pay Structure Outside Education

- Labor market differentially rewards skills and productivity
- Important “recent” changes under the surface
  - Many occupations once closed off to women and minorities no longer are
  - Returns to college quality and technical college skills (degree major) have increased
    - There is an increasing return to graduating from a top college or university (Brewer et al., 1999)
    - There is an increase in the gap (in entry-level salaries) between education and technical majors (Grogger & Eide, 1995)