Teacher Quality: Measurement and Policy

Prof. Jesse Rothstein

Economics 196

Fall 2011

Outline

- Some background
- Measuring teacher quality: “Value added models”
- VA modeling as a search for causal effects
- Evidence on VA models
- Connecting the evidence to policy

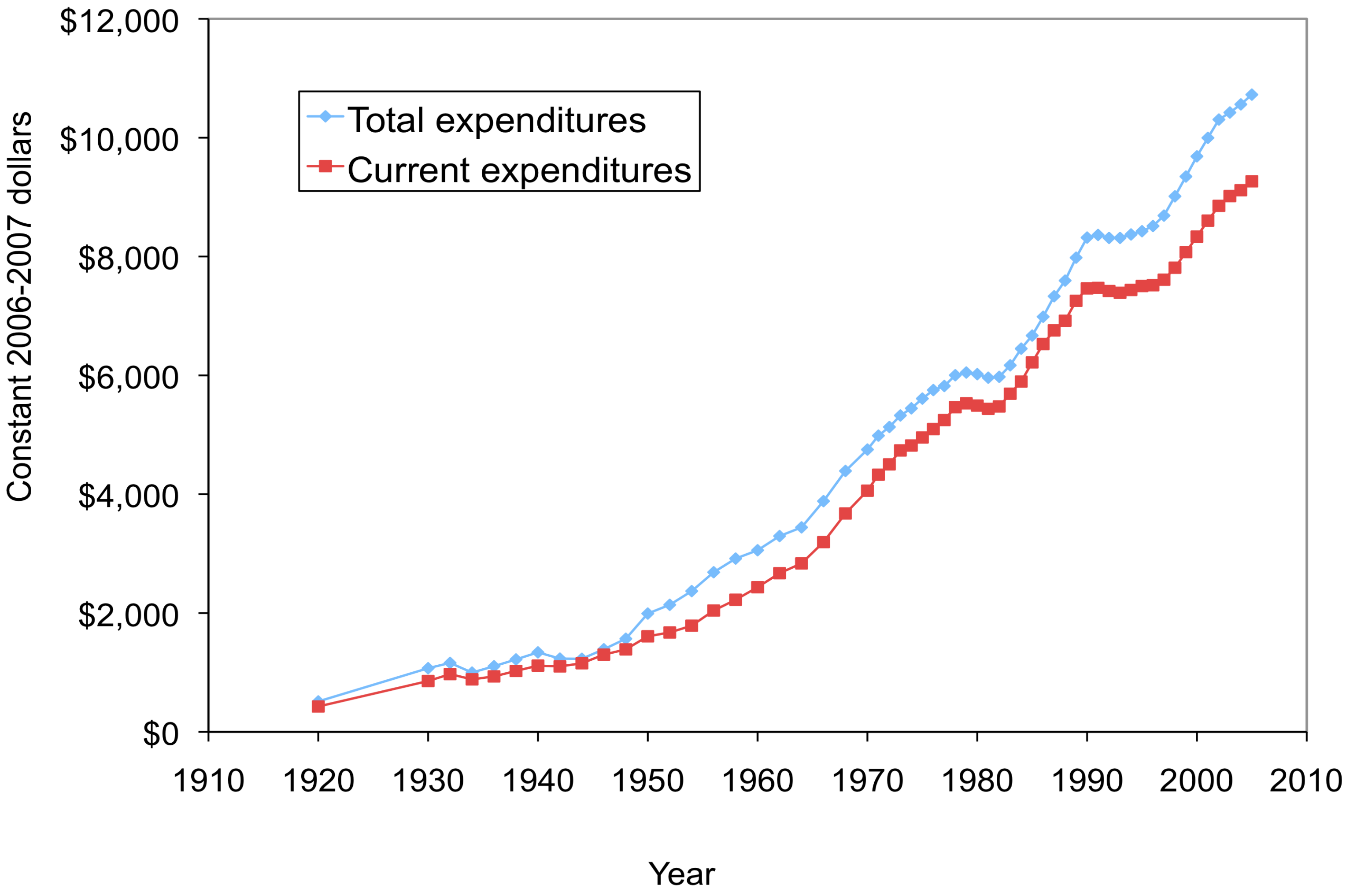
Background

- Rising spending
- Stagnant achievement
- Important, persistent gaps
- U.S. falling behind

Background on U.S. Education

- Rising spending
- Stagnant achievement
- Important, persistent gaps
- U.S. falling behind

Total Expenditure per Pupil in Fall Enrollment

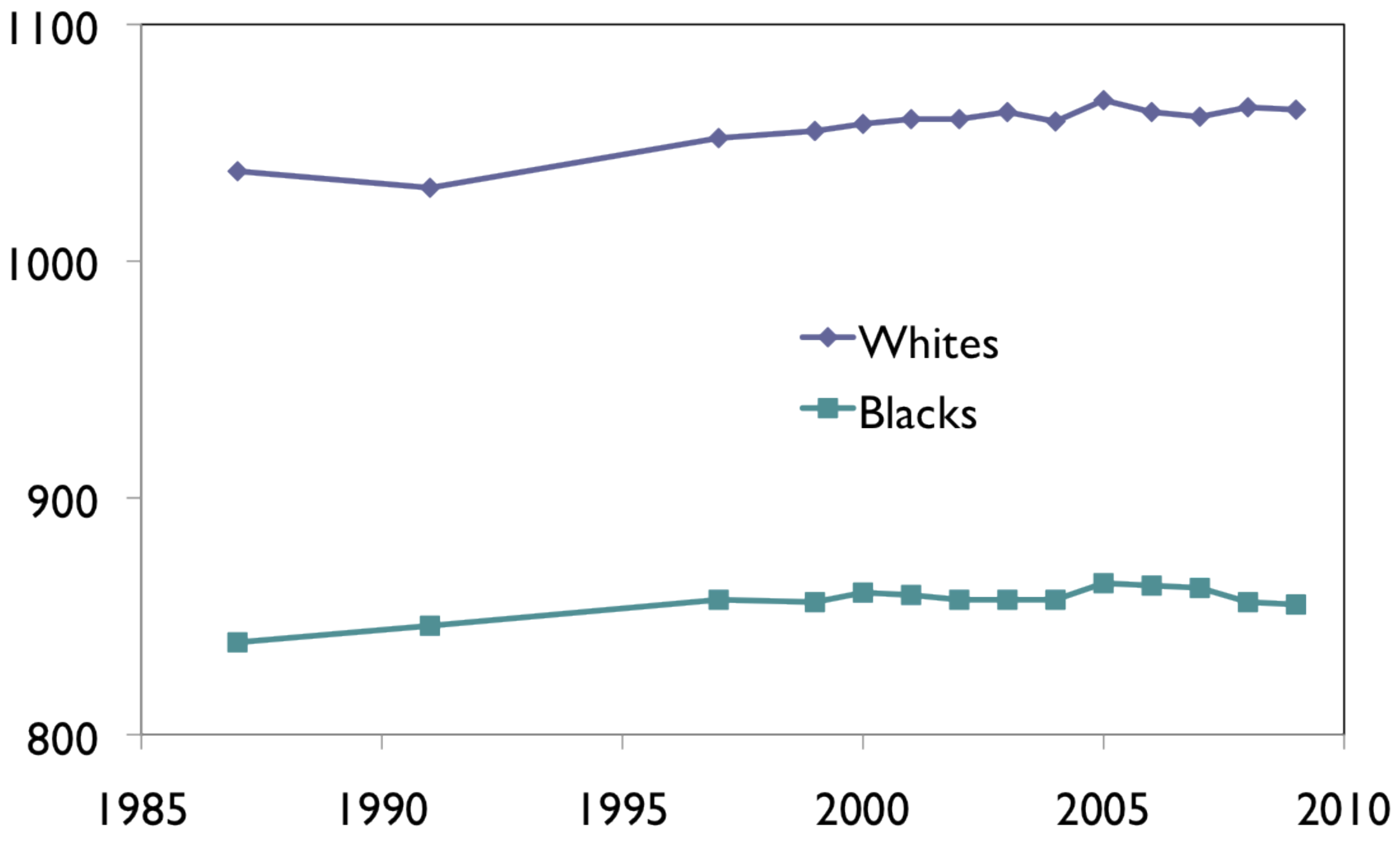


Background

- Rising spending
- Stagnant achievement
- Important, persistent gaps
- U.S. falling behind

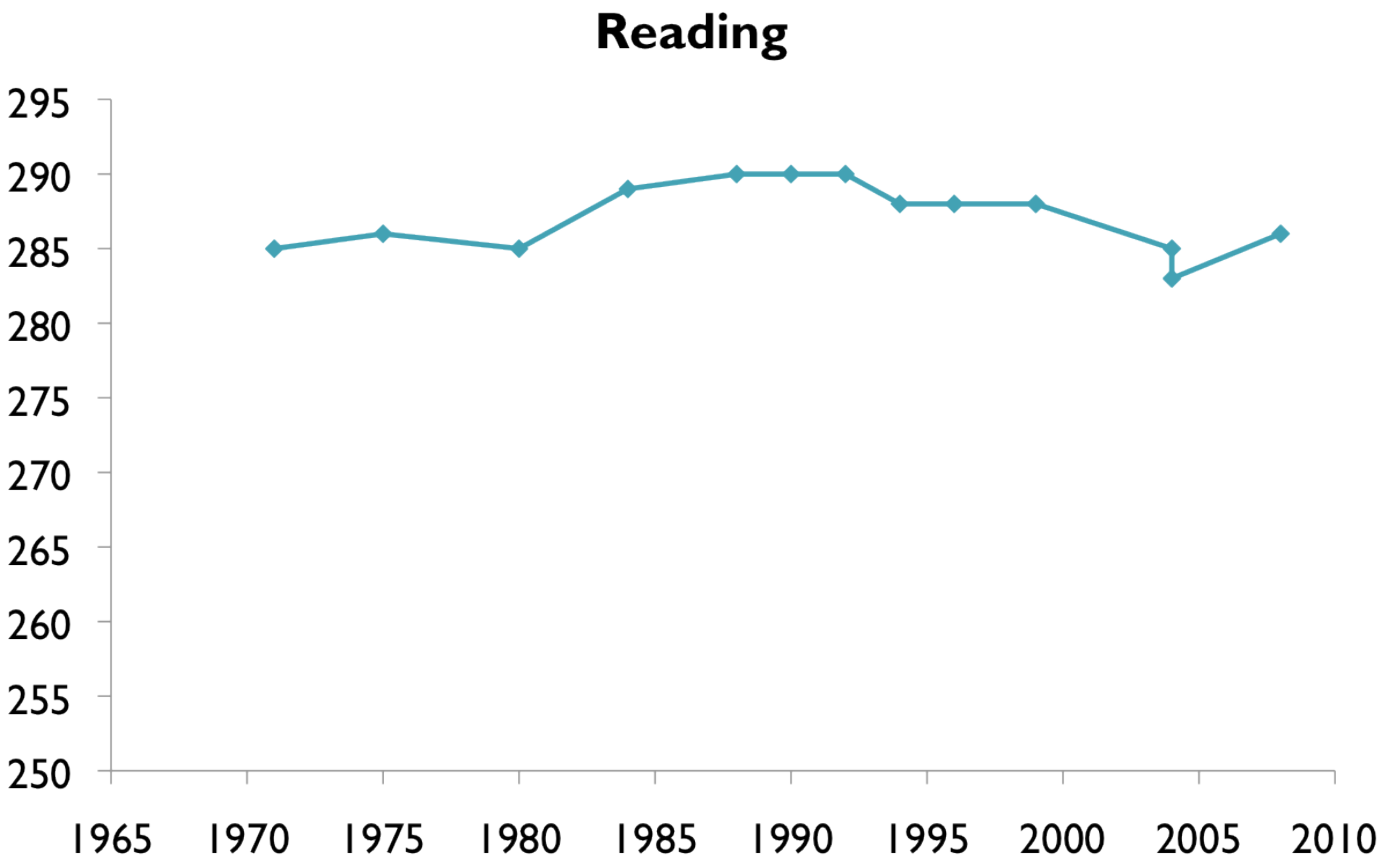
A portrait of stagnation

Average math + verbal SAT scores



A portrait of stagnation

Average 17-year-old reading scores on the NAEP "long-term trend"

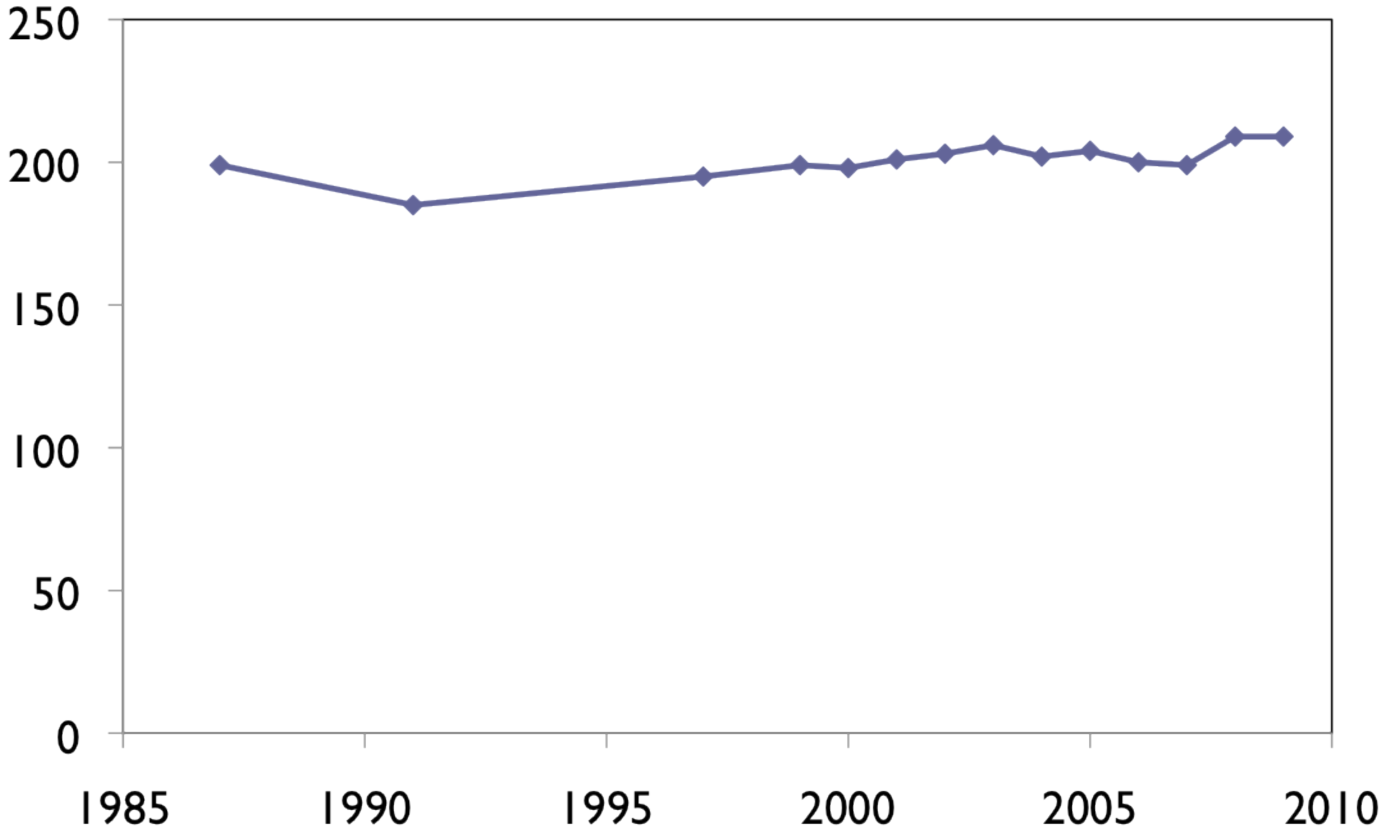


Background

- Rising spending
- Stagnant achievement
- Important, persistent gaps
- U.S. falling behind

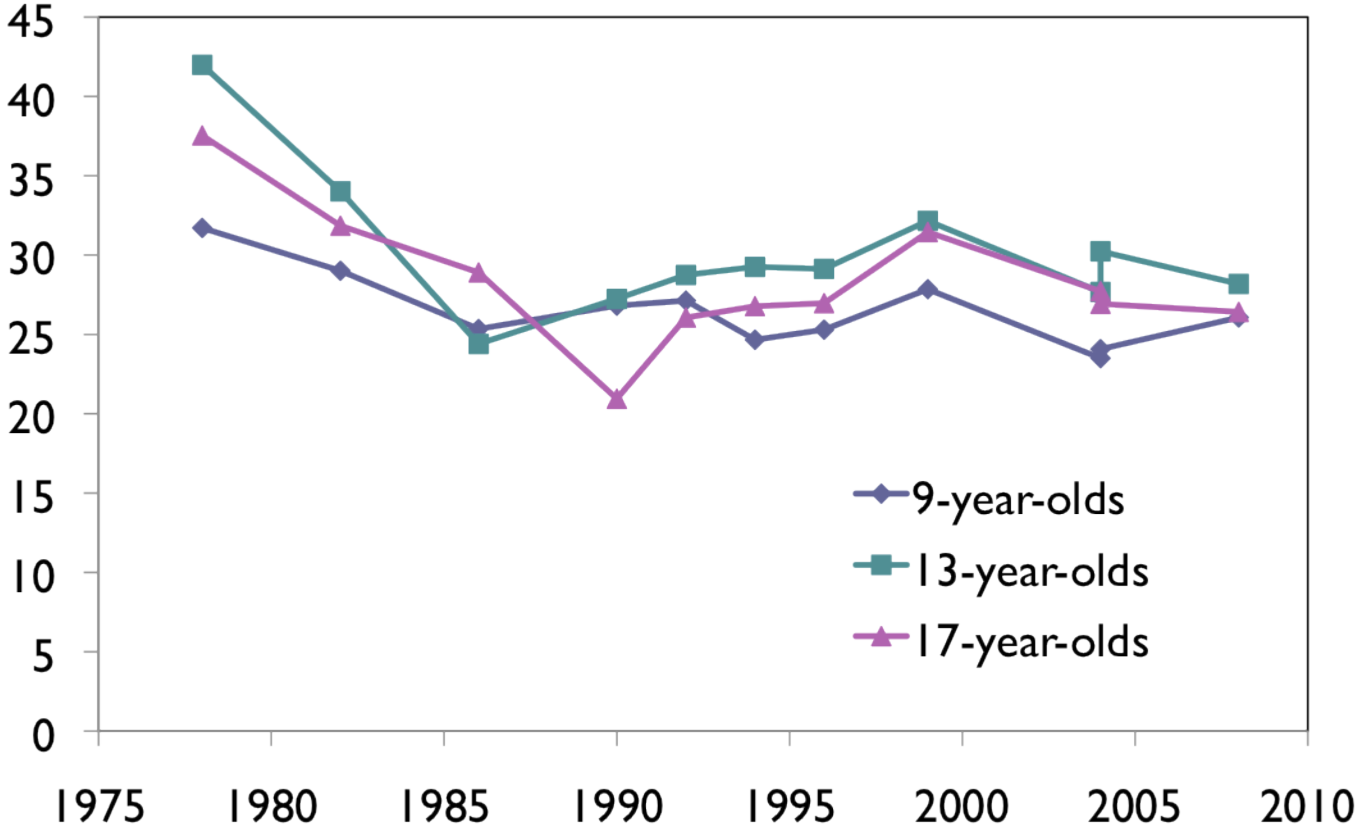
A portrait of stagnation

B-W gap in avg. math + verbal SAT scores



Or maybe not?

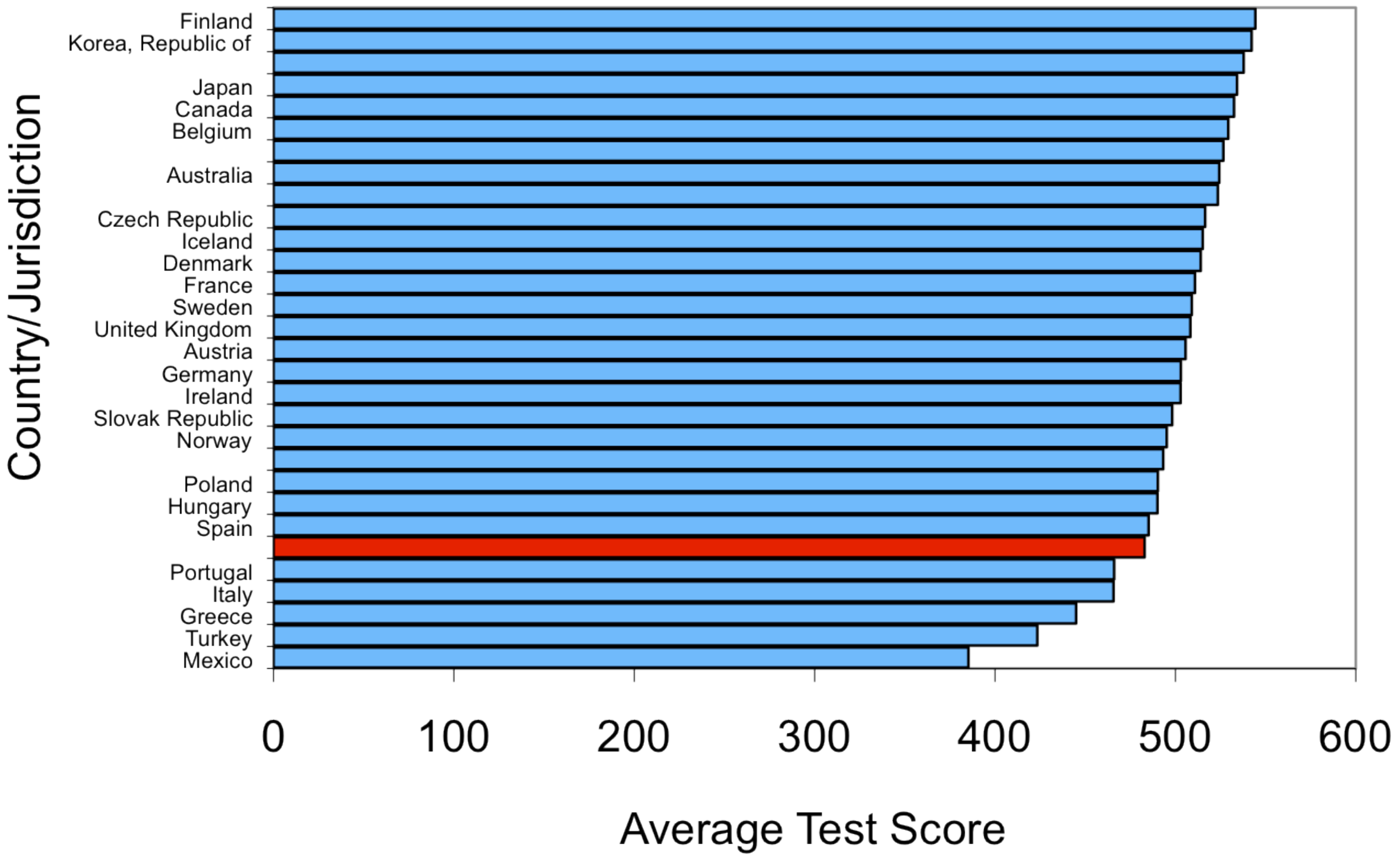
B-W gap in average math scores, by age NAEP Long-Term Trend



Background

- Rising spending
- Stagnant achievement
- Important, persistent gaps
- U.S. falling behind

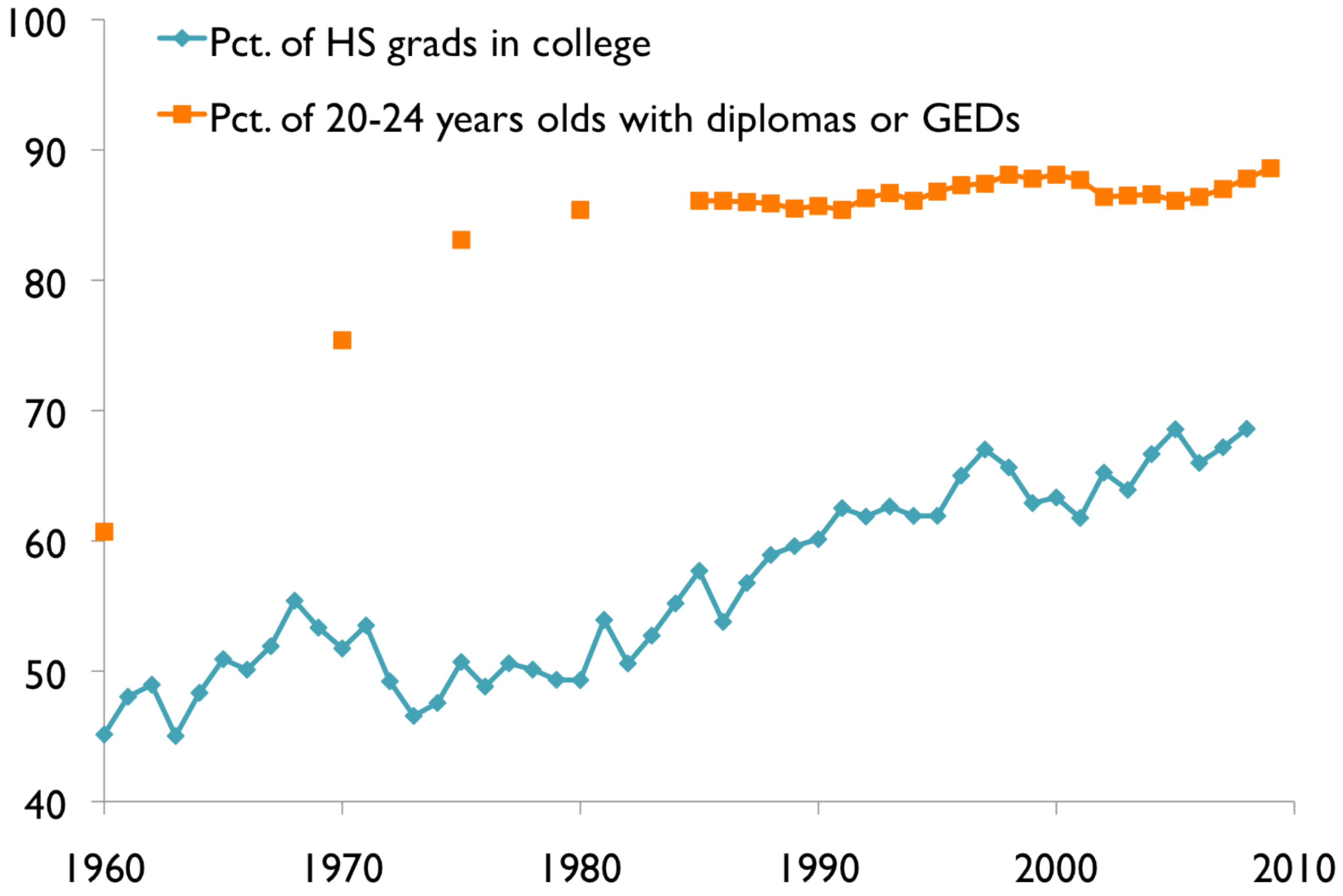
Average Mathematics Proficiency Scores OECD Countries



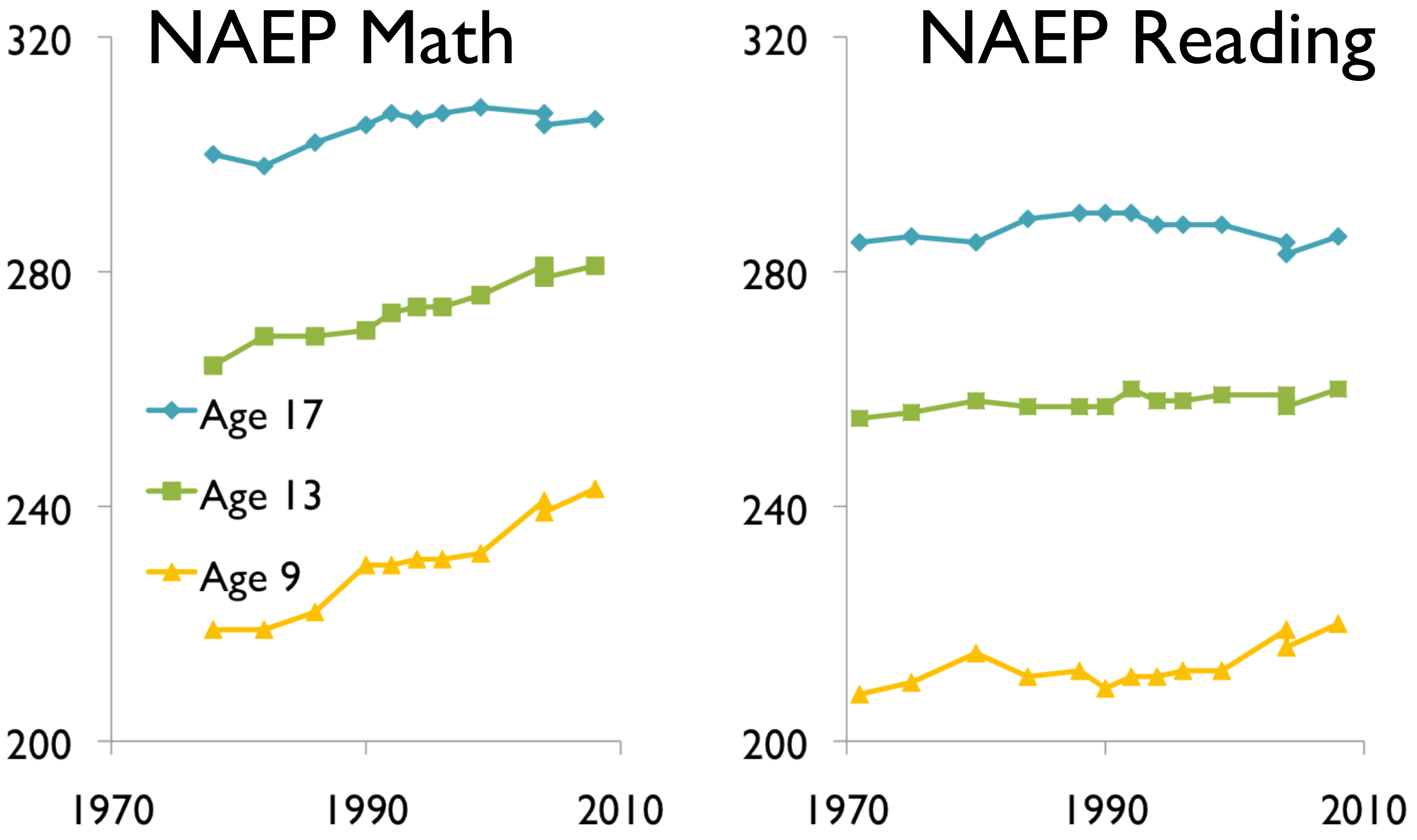
Aside

- Not all of this evidence is airtight
  - Sample selection & 17-year-olds
- Some are cherry-picked
- And some interpretations are tendentious
  - Should special ed spending raise SAT scores?

17-year-olds can be misleading!

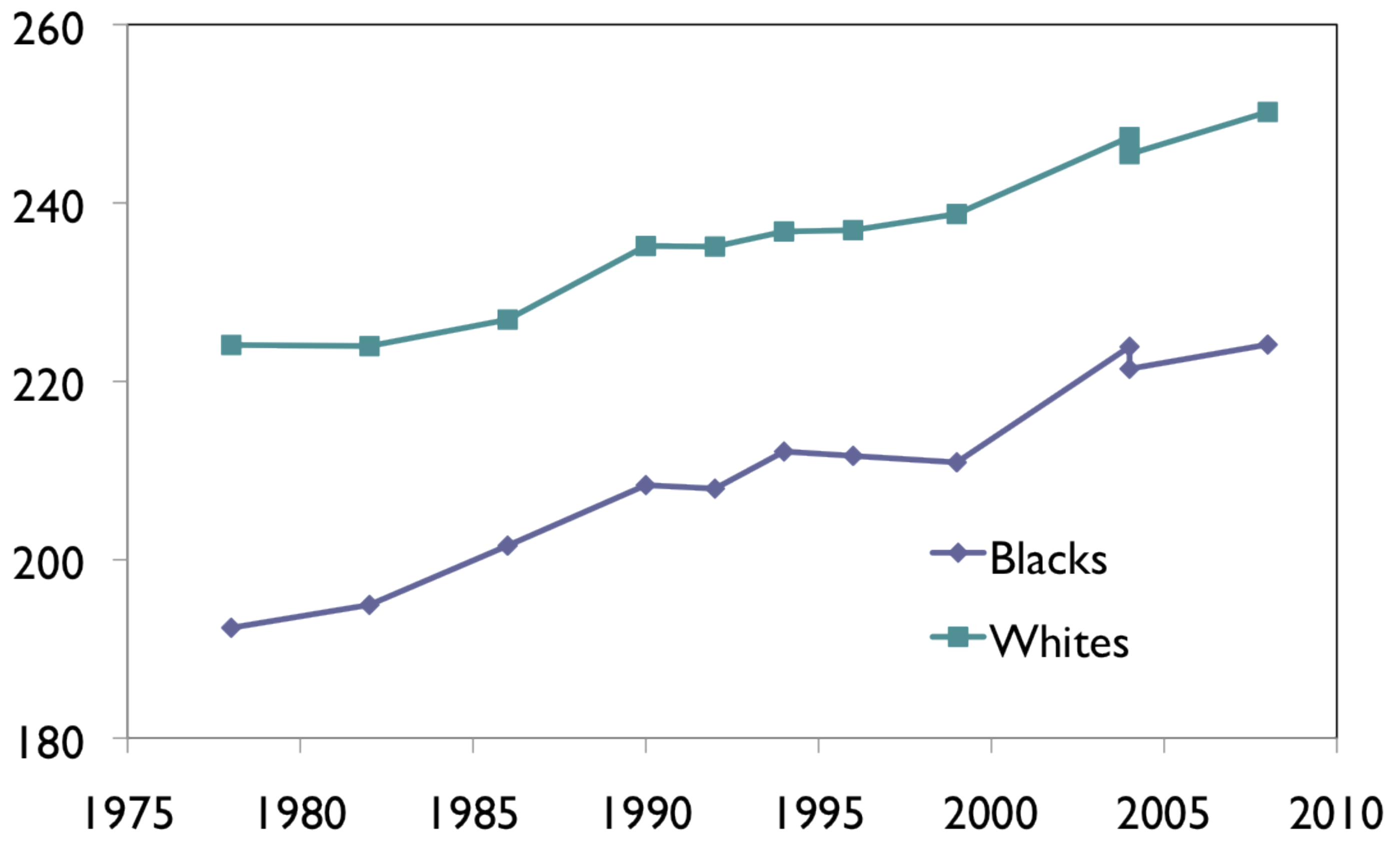


A fuller picture – scores across ages and subjects



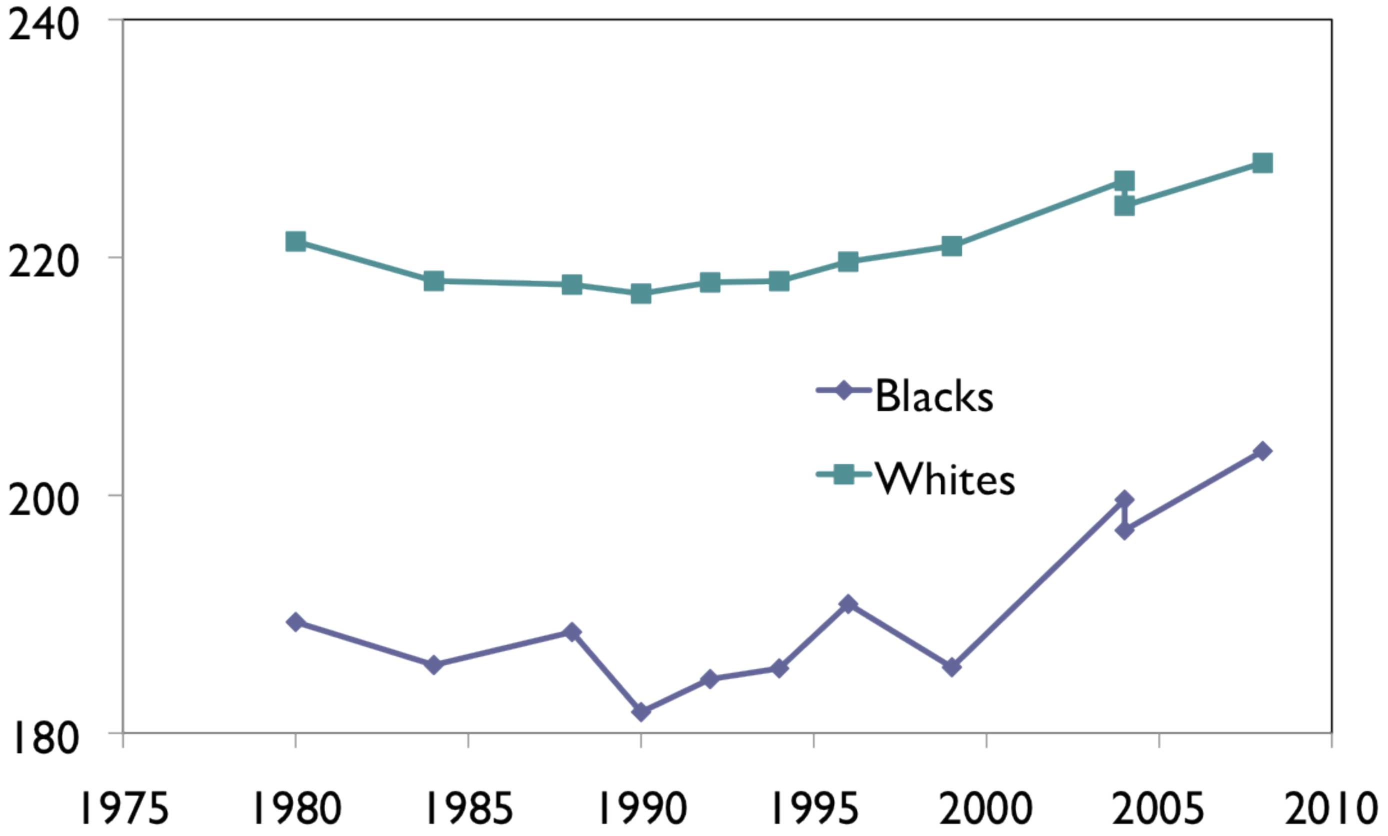
Not exactly a picture of stagnation

Black and white average math scores, age 9 NAEP Long-Term Trend

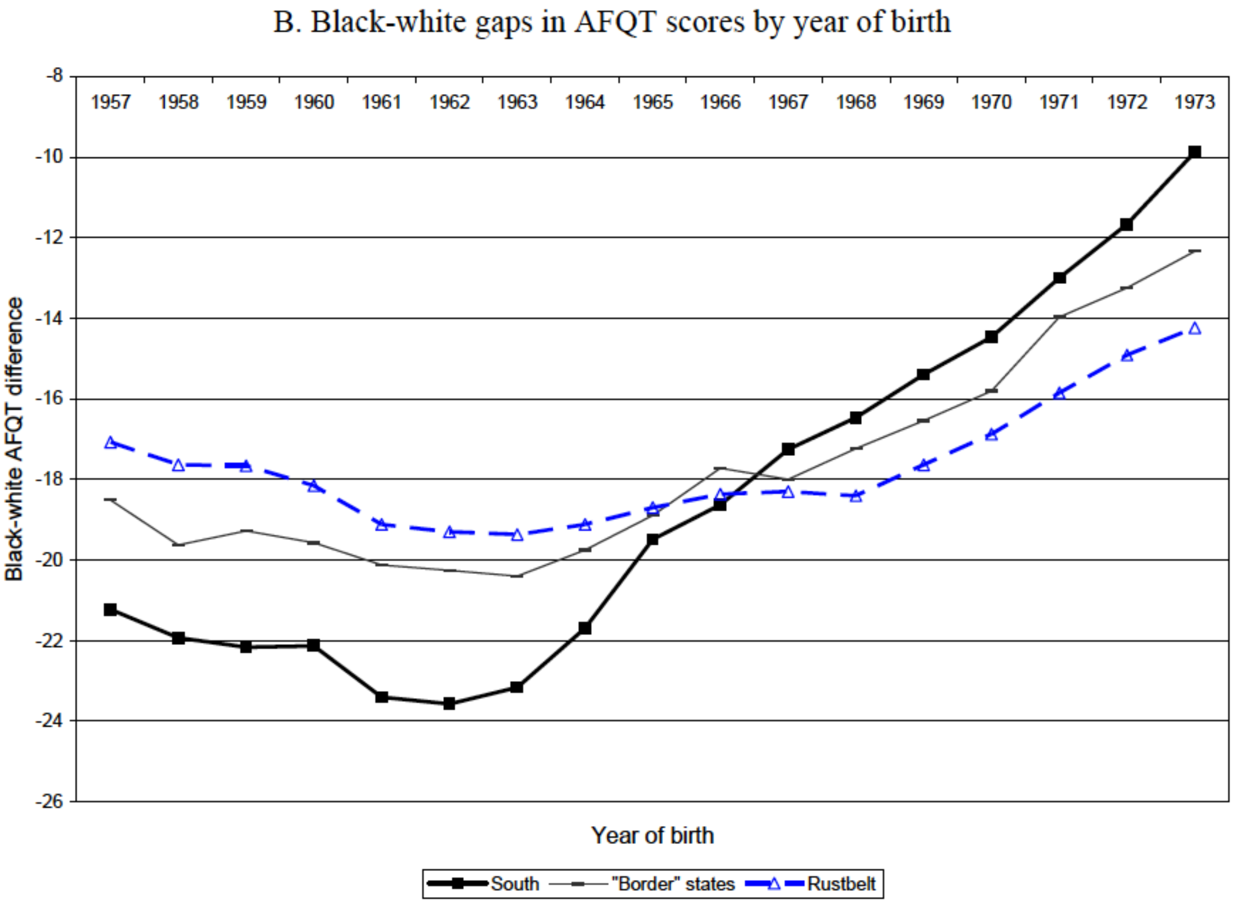


Somewhat more ambiguous

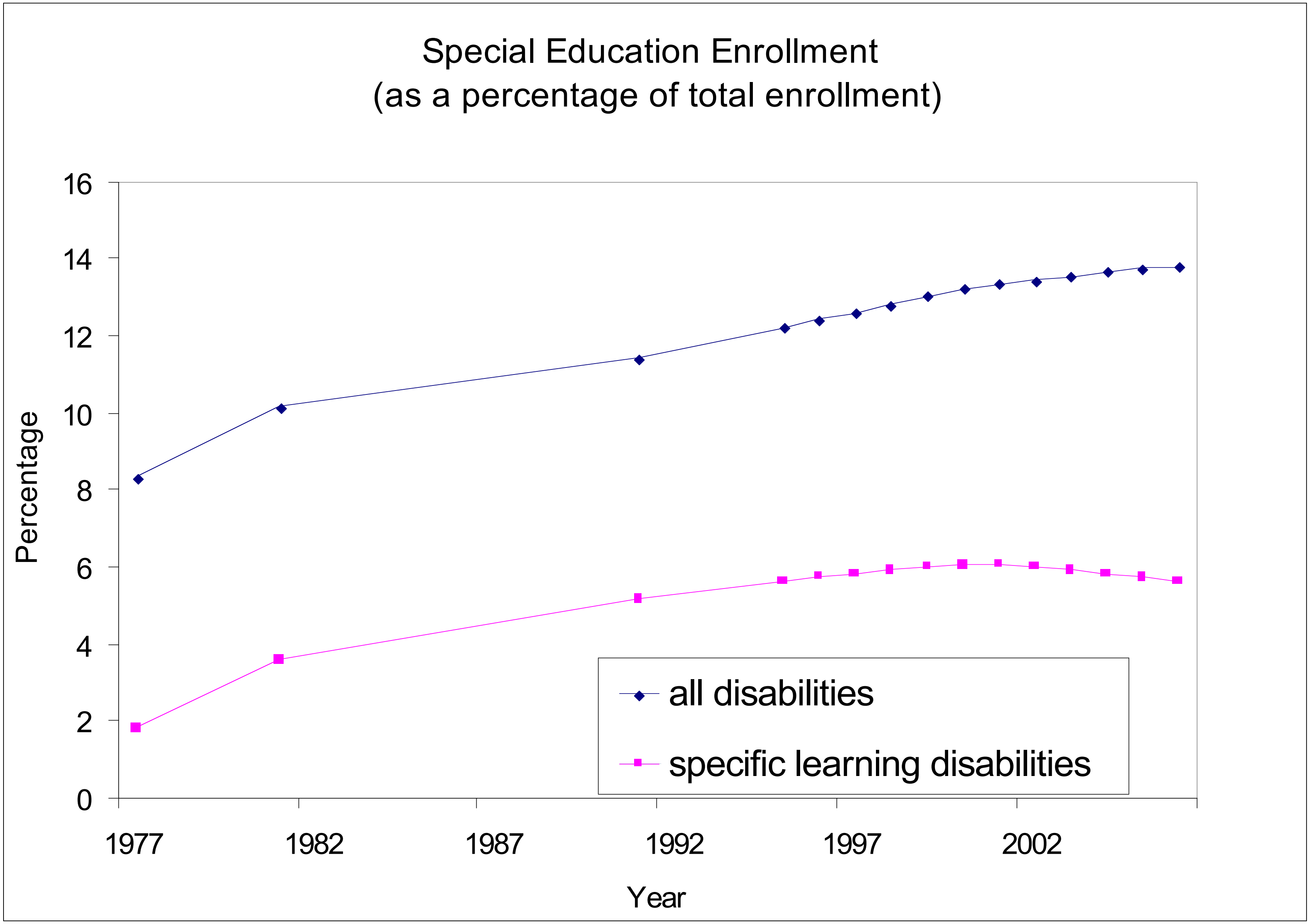
Black and white average reading scores, age 9 NAEP Long-Term Trend



Longer term



Source: Chay, Guryan, and Mazumder, 2009


Special Education Enrollment
(as a percentage of total enrollment)
Percentage
16
14
12
10
8
6
4
2
0
1977
1982
1987
1992
1997
2002
Year
all disabilities
specific learning disabilities

The policy background: A series of failed promises

- Desegregation
- More funds
- More equal funding
- Vouchers
- Charters

→Each came with a lot of fanfare, but little evidence. Most did not change the world.

→Today’s version: Teacher quality

A new consensus: Teachers matter

- "the single most important factor determining whether students succeed in school is not the color of their skin or their ZIP code or even their parents' income – it is the quality of their teacher"

  — *Education Manifesto, signed by Joel Klein, Michelle Rhee, and 14 other superintendents, Oct. 2010*

- "We know what works. What's required, then, to get results from any school is no longer a mystery.…[T]he single most important factor in a student's success after their parent is the person standing at the front of the classroom."

  — *President Obama, speech at TechBoston Academy, March 2011*

- "We know that of all the variables under a school's control, the single most decisive factor in student achievement is excellent teaching."

  — *Bill Gates, "How teacher development could revolutionize our schools," Washington Post, Feb. 2011*

Results from research studies

- Having a top-quartile teacher rather than a bottom-quartile teacher four years in a row would be enough to close the black-white test score gap.
- Having an above average teacher for five years running can completely close the average gap between low-income students and others.
- A teacher one standard deviation better than average raises each student's lifetime earnings by $20,000.
- Replacing the bottom five percent of teachers with average teachers would raise the present value of future U.S. GDP by $100 trillion.

None of these derive from *interventions*.

Even so, can qualify the claims

- "the single most important factor determining whether students succeed in school is not the color of their skin or their ZIP code or even their parents' income – it is the quality of their teacher"

  — *Education Manifesto, signed by Joel Klein, Michelle Rhee, and 14 other superintendents, Oct. 2010*

- "We know what works. What's required, then, to get results from any school is no longer a mystery.…[T]he single most important factor in a student's success after their parent is the person standing at the front of the classroom."

  — *President Obama, speech at TechBoston Academy, March 2011*

- "We know that of all the variables under a school's control, the single most decisive factor in student achievement is excellent teaching."

  — *Bill Gates, "How teacher development could revolutionize our schools," Washington Post, Feb. 2011*

Even so, can qualify the claims

- "the single most important factor determining whether students succeed in school is not the color of their skin or their ZIP code or even their parents' income – it is the quality of their teacher"

  — *Education Manifesto, signed by Joel Klein, Michelle Rhee, and 14 other superintendents, Oct. 2010*

- "We know what works. What's required, then, to get results from any school is no longer a mystery.…[T]he single most important factor in a student's success after their parent is the person standing at the front of the classroom."

  — *President Obama, speech at TechBoston Academy, March 2011*

- "We know that of all the variables under a school's control, the single most decisive factor in student achievement is excellent teaching."

  — *Bill Gates, "How teacher development could revolutionize our schools," Washington Post, Feb. 2011*

Even so, can qualify the claims

- "the single most important factor determining whether students succeed in school is not the color of their skin or their ZIP code or even their parents' income – it is the quality of their teacher"

  — *Education Manifesto, signed by Joel Klein, Michelle Rhee, and 14 other superintendents, Oct. 2010*

- "We know what works. What's required, then, to get results from any school is no longer a mystery….[T]he single most important factor in a student's success after their parent is the person standing at the front of the classroom."

  — *President Obama, speech at TechBoston Academy, March 2011*

- "We know that of all the variables under a school's control, the single most decisive factor in student achievement is excellent teaching."

  — *Bill Gates, "How teacher development could revolutionize our schools," Washington Post, Feb. 2011*

Why such confidence?
A new technology: "Value Added Models"

- Estimate a teacher's effectiveness based on the average achievement of that teacher's students.
- Effective teaching = unusually large test score increases.
- Generates estimates for all teachers – or at least for the roughly 40% whose students are tested.
- Much cheaper than classroom observations.
- Indicates wide variability in teacher quality.
- But teachers aren't magic: Moving from a 25th percentile teacher to a 75th percentile teacher raises an average student from the 50th percentile to about the 56th.

How to estimate a teacher's value added?

- Options:
  - Average scores of her students
  - Average *gain* scores
  - Average scores, relative to other students who started in the same place.
  - Something else?
- How to choose? What are we trying to accomplish?

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?
- A: A teacher's causal effect on her students.

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?
- A: A teacher's causal effect on her students.
- Q: Relative to what?

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?
- A: A teacher’s causal effect on her students.
- Q: Relative to what?
- A: A “typical” teacher

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?
- A: A teacher's causal effect on her students.
- Q: Relative to what?
- A: A "typical" teacher
- Q: How to recover it?

Value added as a nonexperimental estimator

- Q: What is the estimand of interest?
- A: A teacher’s causal effect on her students.
- Q: Relative to what?
- A: A “typical” teacher
- Q: How to recover it?
- A: By analogy to a random assignment experiment.

Value added with random assignment

- Imagine students are randomly assigned to teachers.
- How can we estimate the teacher's causal effect? (relative to what?)

Value added with random assignment

- Imagine students are randomly assigned to teachers.
- How can we estimate the teacher’s causal effect? (relative to what?)
- Just about any comparison will do!
  - Average scores
  - Average gain scores
  - Average scores, controlling for previous scores

Value added without random assignment

- Imagine students are *not* randomly assigned to teachers.
- How can we estimate the teacher's causal effect?
- Which of our options still work?

The Fundamental Problem of Causal Inference

- Imagine we want to know the effect of having Ms. Jones vs. Ms. Smith.
- For students who get Ms. Jones, we don’t know what their scores would have been had they gotten Ms. Smith.
- For students who get Ms. Smith, we don’t know what their scores would have been had they gotten Ms. Jones.
- Randomization solves the FPCI!

The Fundamental Problem of Causal Inference without random assignment

- If we don't have RA, need some assumptions (actually, we need them even with RA).
- What would have happened to Ms. Jones' students had they had Ms. Smith?
  - Would have been the same as Ms. Smith's students → as good as RA.
  - Would have been the same as those of Ms. Smith's students with the same previous scores → Value added model!

A simple VA model

$$y_{it} = \alpha + y_{i,t-1}\beta + T_{it}\gamma + X_{it}\theta + \epsilon_{it}$$

y_{it}	Test score of student i in year t.
$y_{i,t-1}$	Same student’s score the previous year.
T_{it}	=1 if Ms. Jones, 0 if Ms. Smith
γ	The effect of having Ms. Jones
X_{it}	Control variables (e.g., race, class size)

- Lots of variations, but same basic idea.
- Identifying assumption: $cov\left(T_{it}, \epsilon_{it}\right) = 0$

- What is the “counterfactual” assumption?

What would violate the identifying assumption?

- If characteristics that predict future achievement also predict classroom assignments.
- That is, any kind of tracking based on variables that we can't observe / don't control for.
- How to assess?
  - Find some of those variables, and see if they predict both y and T.
  - Alternatively, see if T predicts those variables even though it couldn't possibly cause them.
  - Candidates: Anything observed by actors (principal, teachers, parents) before classroom assignments are made.

→ Look to see if 5th grade teachers (appear to) affect 3rd grade test scores.

A simple falsification test

$$Z_i = \alpha^Z + y_{i,t-1}\beta^Z + T_{it}\gamma^Z + X_{it}\theta^Z + \epsilon_{it}^Z$$

- Choose a Z for which we know the causal effect of T_{it} is zero.
- Any will do. But best Z is one that we think predicts y_{it}.
- Examples:
  - Height
  - $Y_{i,t-2}$

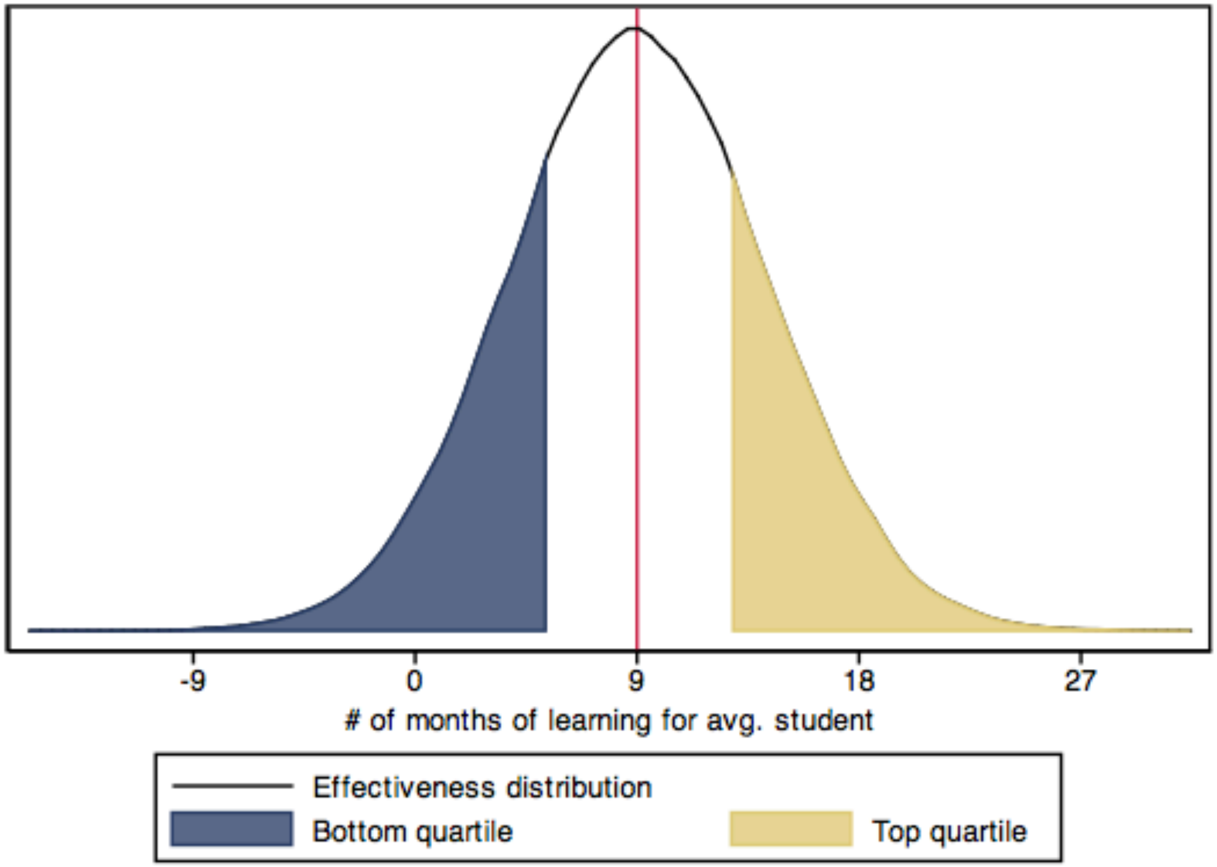
Guess what?

	5th grade math scores	3rd grade math scores	5th grade reading scores	3rd grade reading scores
SD of 5th grade teacher effect on:	0.150	0.067	0.109	0.076

What does this mean?
-Systematic variation in students’ 3rd grade scores, controlling for 4th grade scores.
-Sorting, which VA models attribute to teacher effectiveness.

What’s more:
-Similar for other VA specifications
-Averaging across multiple years doesn’t solve the problem.
-Neither does controlling for full test score history: 5th grade teachers “affect” 4th grade TV watching.

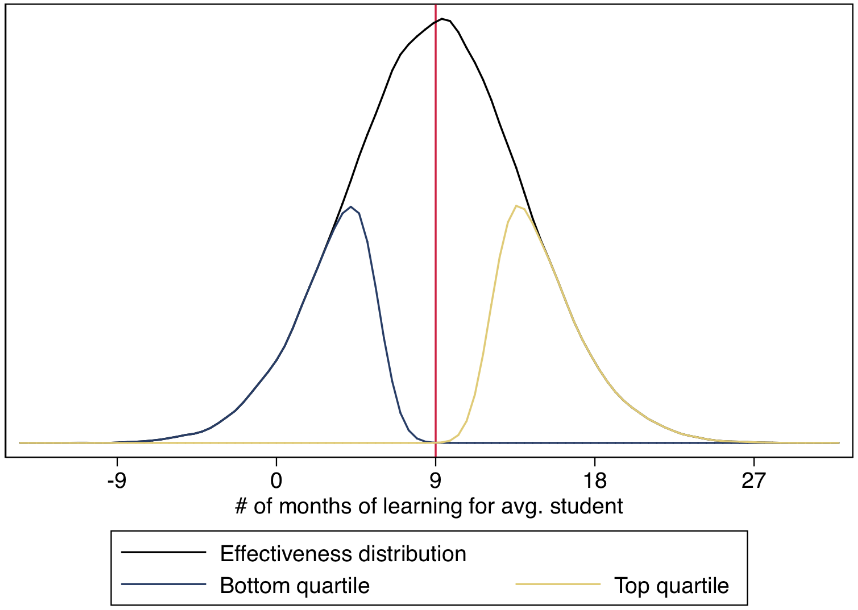
Value-added estimates of teacher effectiveness: An illustration

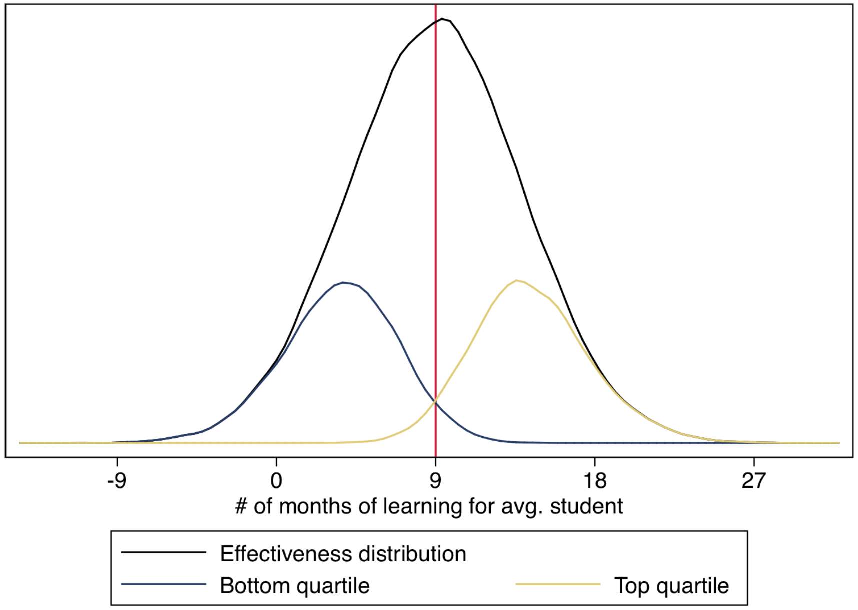


Misclassification due to nonrandom assignment

Best case
(selection on observables)

Not quite the worst case
(selection on true achievement +
observed scores)



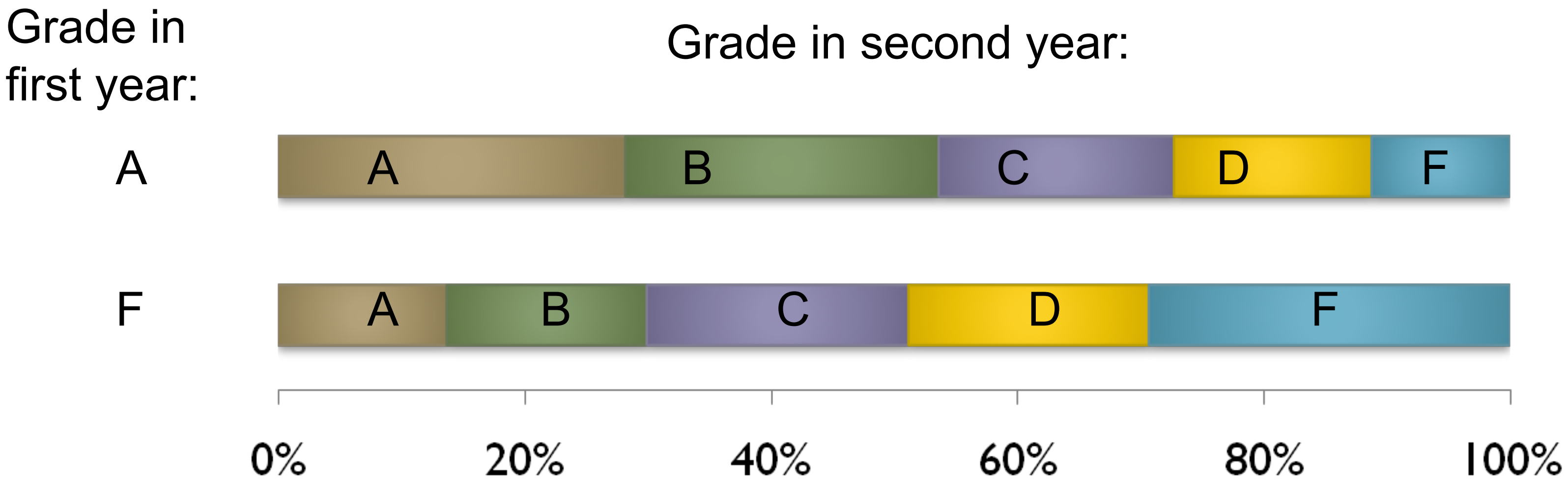


Other recent research results

- Many teachers indicated as effective for one class are ineffective for others (and vice versa).
- Many teachers effective for one test are ineffective for another (and vice versa).
- Many teachers effective in the short-run are ineffective for long-run outcomes (and vice versa).

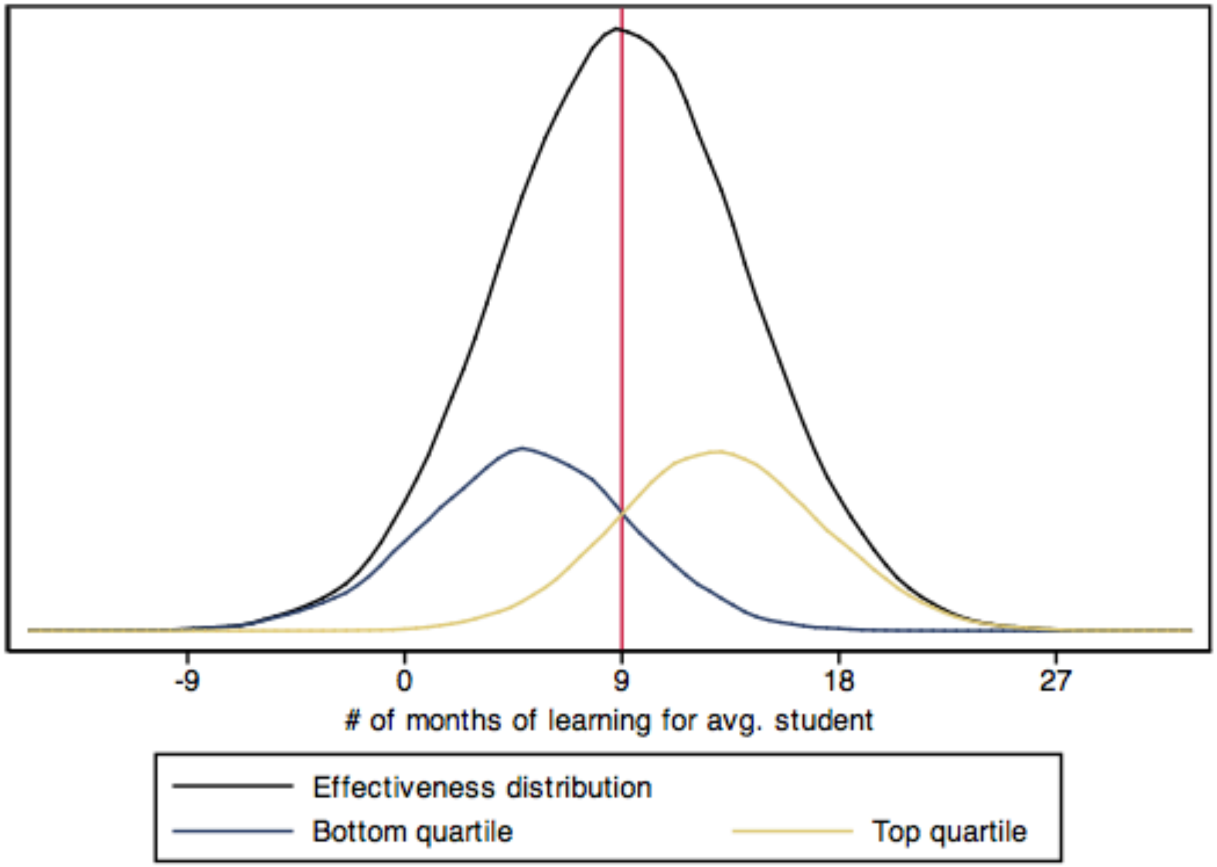
Value-added estimates are extremely noisy.

- Only 20-30 students in average class (even in CA!)
- Consider classification of teachers into 5 categories (A-F) in two consecutive years.

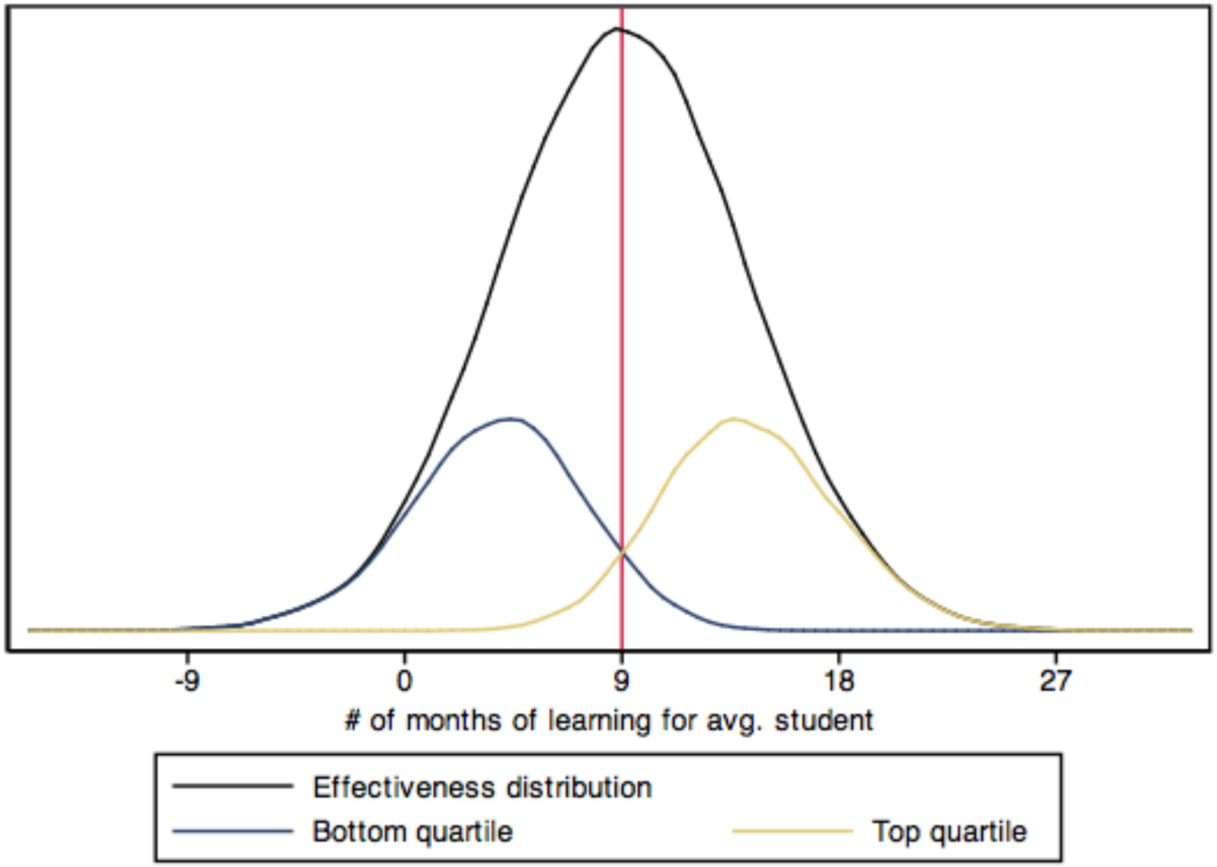


Average across 5 Florida districts. Grades A-F correspond to quintiles 1-5. Source: Sass (2008).

Value-added estimates of teacher effectiveness: *Misclassification due to noisy estimates*



Value-added estimates of teacher effectiveness: *Misclassification with three years of data*

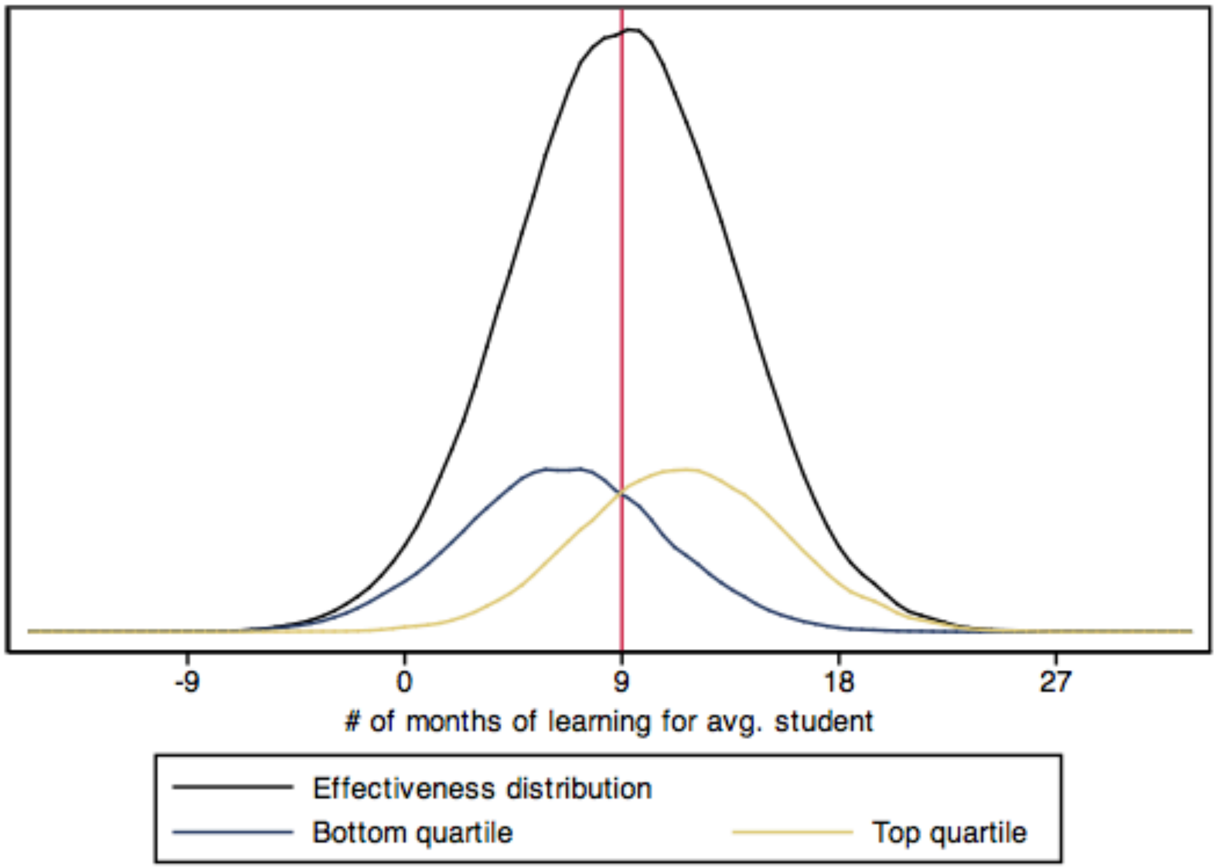


Even setting noise aside, what are VA models measuring?

VA models assume teacher effectiveness has one dimension. If so, any (decent) test should give similar estimates. But:

1. 20%-30% of teachers in top quartile in terms of impacts on state assessment scores are in bottom half of impacts on more conceptually demanding tests (and vice versa).
2. Teachers' estimated effectiveness is very different for "Procedures" and "Problem Solving" subscales of the same math test.
3. Teacher effects on high-stakes tests are only slightly related to effects on low stakes tests, and dissipate more quickly.

Value-added estimates of teacher effectiveness: *Which test is the right one?*



And that's not all…
More concerns about VA models:

1. Models don't distinguish teacher contributions from differences among students.
2. Rewards only learning measured by the test – not other subjects or other skills.
3. May capture "teaching to the test" rather than real effectiveness.
4. "Fadeout" isn't well understood.
5. Many teachers aren't covered.

Campbell's Law

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

"[A]chievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. **But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.**"

→ Whatever the problems are with VA now, they'll get worse if we raise the stakes.

Campbell's Law in Practice

- "I'm scared to teach in the 4th grade. I'm scared I might lose my job if I teach in an [ELL] transition grade level, because I'm scared my scores are going to drop, and I'm going to get fired because there's probably going to be no growth."
- "When they say nobody wants to do 4th grade – nobody wants to do 4th grade! Nobody."
- "I found out that I [have been] competing with myself."
- "Every year I have the highest test scores, I have fellow teachers that come up to me when they get their bonuses…One recently came up to me [and] literally cried - 'I'm so sorry.'… I'm like, don't be sorry…It's not your fault. Here I am… with the highest test scores and I'm getting $0 in bonuses. It makes no sense year to year how this works…. How do I, how do I… you know… I don't know what to do. I don't know how to get higher than a 100%."
- "I have students [in a 5th grade gifted reading class] who score at the 6th 7th 8th-grade levels in reading. But I'm like please babies, score at the 9th grade level, cause if you don't score at the 9th or 10th grade or higher in 5th grade with me, I'm going to show negative growth. Even though you, you're gifted and you're talented, and you're high! I can only push you so much higher when you are already so high. I'm scared."

THE DESIGN OF TEACHER QUALITY POLICY

Where do we stand now?

- Reform movement has lots of momentum
  - Race To the Top
  - Blueprint for revised No Child Left Behind / Elementary and Secondary Education Act
  - Teacher Incentive Fund
  - Support from Gates, Broad, other powerful outsiders
  - Recent events in Wisconsin
- So far, a high ratio of rhetoric to results
- What would a serious teacher quality policy look like?

Two routes to higher teacher quality

1. Induce teachers to work harder / better.
2. Induce better people to enter and remain in the teaching profession.

- Barrier: The salary & retention schedule
  - "Tenure" after 2-3 years is typical.
  - Pay depends solely on education (number of graduate credits) and experience.
- Proposals: Performance-based compensation and retention
  - Retain only high-quality teachers
  - Differentiated pay to reward quality

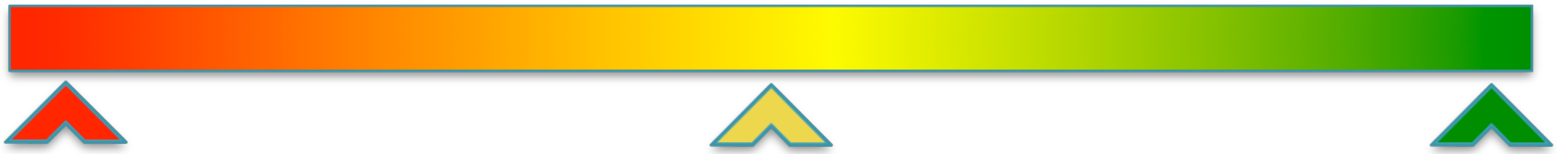
Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers		
Goal 2: Higher ability teachers in the profession		

Assessment: Two goals and two policies

	Policy 1: **Differentiated pay**	**Policy 2:** **High-stakes retention decisions**
Goal 1: Better output from existing teachers		
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

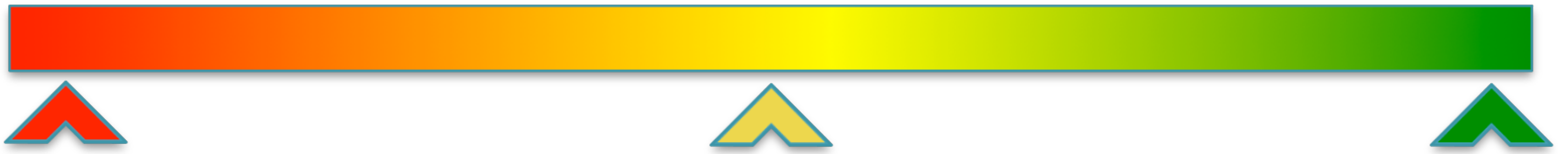
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation.	
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

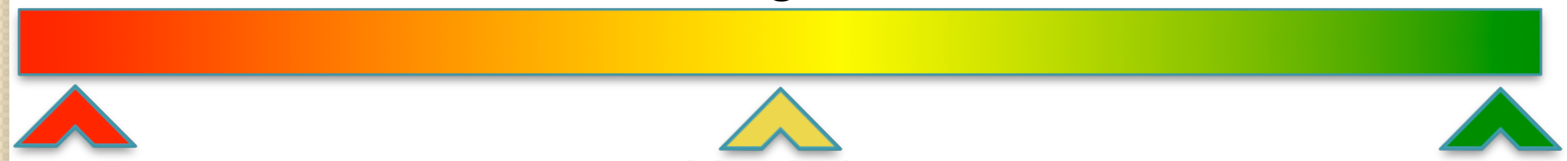
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	**Policy 2: High-stakes retention decisions**
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation.	
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

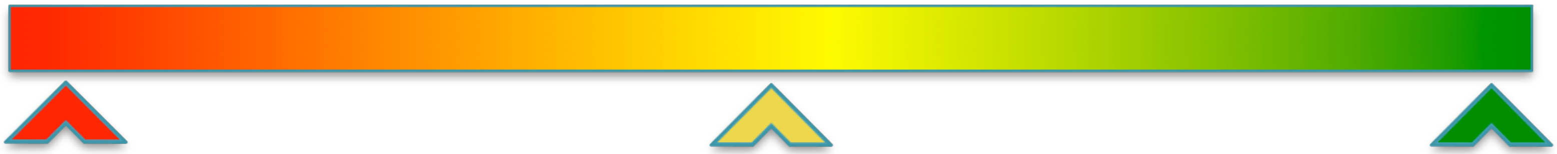
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

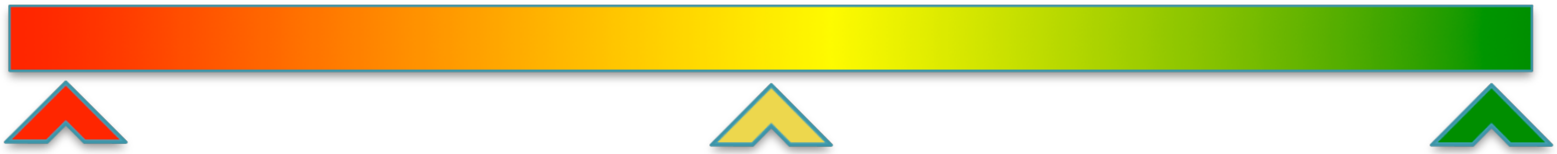
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

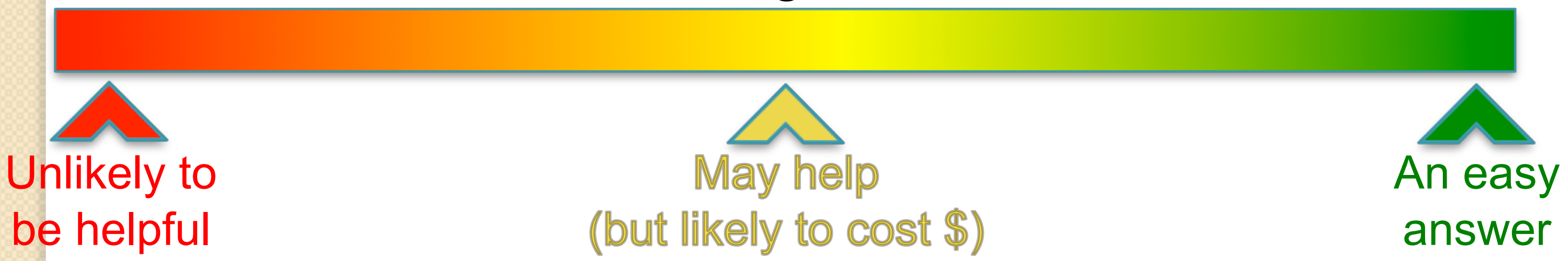
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession		

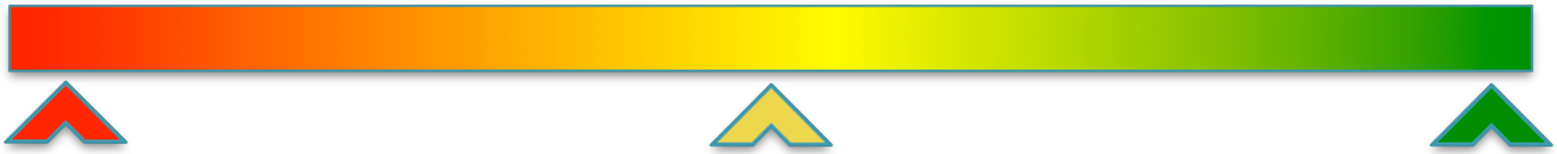
Scoring rubric



Assessment: Two goals and two policies

	Policy 1: **Differentiated pay**	**Policy 2:** **High-stakes retention decisions**
Goal 1: Better output from existing teachers	🟥 Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	🟥 Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession		

Scoring rubric



Unlikely to be helpful

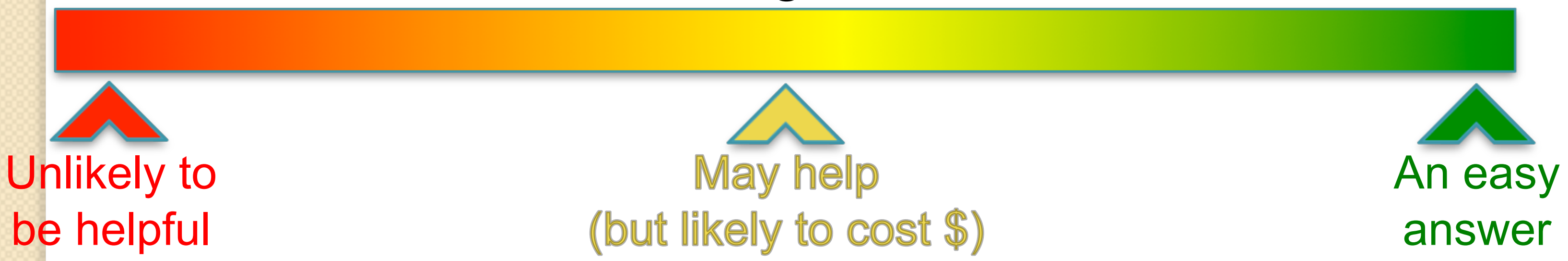
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: **Differentiated pay**	**Policy 2:** **High-stakes retention decisions**
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession	Attracts teachers who expect to win the competition.	

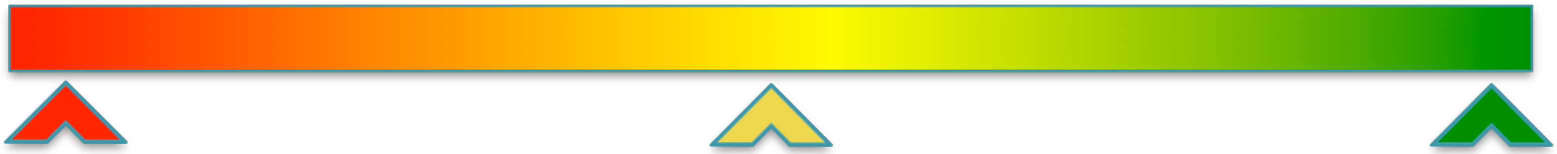
Scoring rubric



Assessment: Two goals and two policies

	Policy 1: **Differentiated pay**	**Policy 2:** **High-stakes retention decisions**
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession	Attracts teachers who expect to win the competition.	

Scoring rubric



Unlikely to be helpful

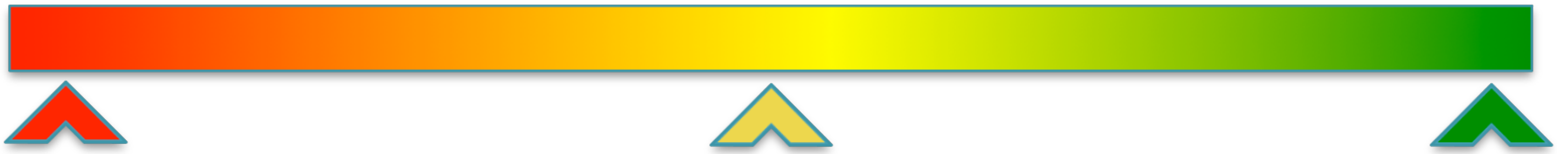
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: Differentiated pay	Policy 2: High-stakes retention decisions
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession	Attracts teachers who expect to win the competition.	Weeds out worst teachers (but they need to be replaced!)

Scoring rubric



Unlikely to be helpful

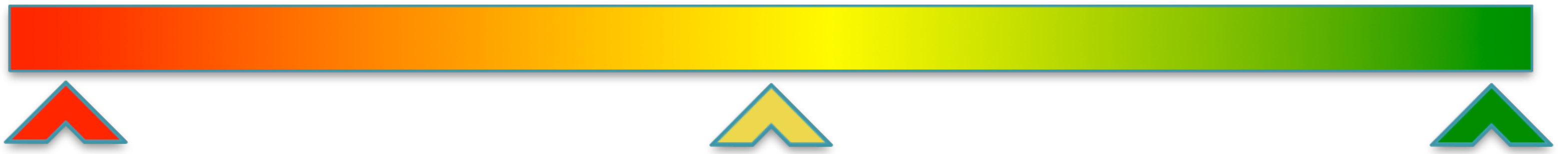
May help (but likely to cost $)

An easy answer

Assessment: Two goals and two policies

	Policy 1: **Differentiated pay**	**Policy 2:** **High-stakes retention decisions**
Goal 1: Better output from existing teachers	Incentive for all teachers to work harder. May undermine cooperation. But experimental results bad.	Scare new teachers into working harder. No effect on later career teachers.
Goal 2: Higher ability teachers in the profession	Attracts teachers who expect to win the competition.	Weeds out worst teachers (but they need to be replaced!)

Scoring rubric



Unlikely to be helpful

May help (but likely to cost $)

An easy answer

What's the alternative? Lessons from other skilled professions.

- Mechanical incentive pay is extremely rare.
- Evaluations are subjective, conducted by highly-trained and highly-skilled managers.
- High ratio of managers to managed.
- Stakes aren't too high.
- Evaluations are formative, not just summative, with real effort to help people improve.

→ Improving teacher quality will be a long, expensive slog, not a panacea.