



Value-Added Measures in Education: Part I

Douglas N. Harris

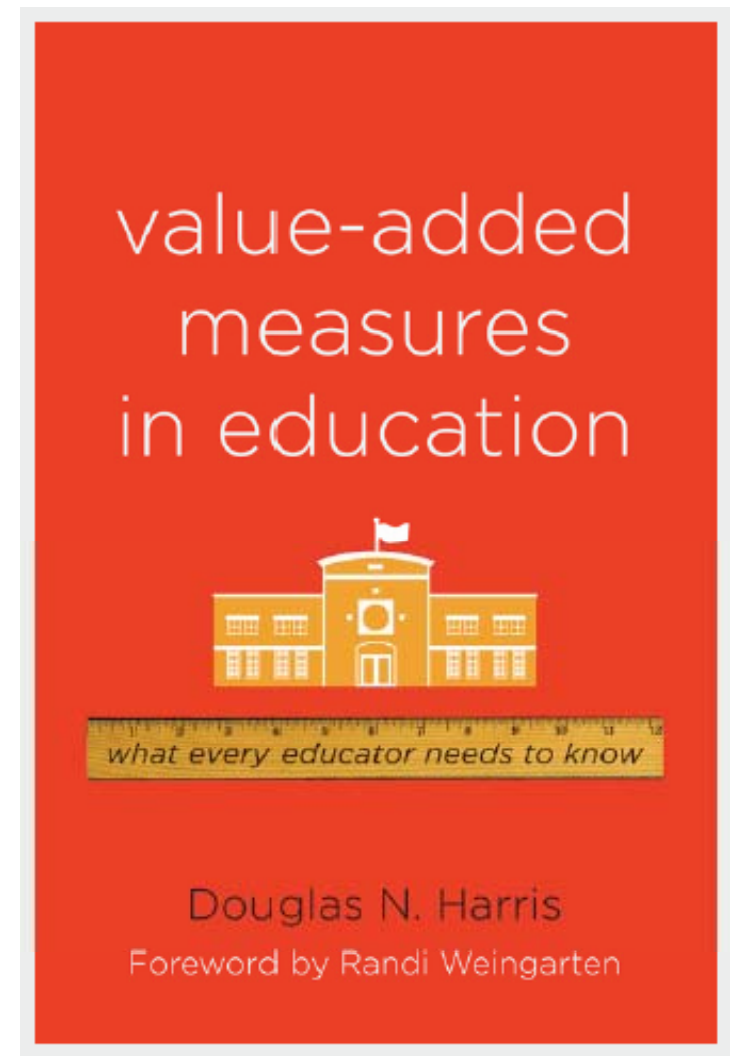
Associate Professor of Educational Policy and Public Affairs
University of Wisconsin at Madison

April 27, 2011

State of Missouri

Based on the book

- Not a technical manual, a user's manual
- Educators, as the targets of accountability, deserve to understand how they are being evaluated
- Policymakers need to better understand the measures as they embed them in policy



<http://www.hepg.org/main/hep/Index.html>

Chapters 1-4: Arguments for VA


1. Exploring the Potential of Value-Added Performance Measures
2. Using—and Misusing—Achievement Tests
3. Measuring Student Growth
4. Creating Value-Added Measures
5. Understanding Statistical Errors in VA

** We will cover these chapters today (with specific references to the Missouri system)

Chapters 5-11: Cautions & Recommendations

6. Shifting from School to Teacher VA
7. Marshaling the Evidence about VA
8. Evaluating VA Measures and Avoiding Double Standards
9. Using Value-Added to Improve Teaching and Learning
10. Creating Value-Added that Match Their Uses
11. Addressing Key Trade-Offs, Misconceptions, and Questions

** We will cover these on Friday



Chapter 1:
Exploring the Potential of Value-
Added Performance Measures

Guiding Principle

- Hold people accountable for what they can control
 - Part 1: “Hold people accountable . . .”
 - Meaning that accountability is important
 - Part 2: “. . . for what they can control.”
 - Meaning that the details matter

Failure #1: We Measure Teacher Performance Poorly


- We rely on credentials like degrees, certification, and years of experience—which are largely unrelated to teacher performance
 - Not surprising, teaching is complex & relationship-driven
- Formal teacher evaluations are hardly any better—they ignore “the technical core” of teaching
- Wide agreement on the above

One Result is that Teacher Performance Varies (a lot?)

- “New” research suggests that teacher effectiveness varies a great deal, even within individual schools
 - Some even argue that we can eliminate the achievement gap simply by reassigning the most effective teachers to minority children
 - Differences are exaggerated but the larger conclusion about variation is not really in dispute
 - Educators think we’ve known this for a long time

Failure #2: Snapshot Problem

- Snapshot = Any measure of student outcomes at a single point in time
- The Snapshot Problem: Students enter the classroom at very different starting points, because of factors outside the control of the school
 - The “starting gate inequality”: students from high-income families start kindergarten with test scores 60% higher than low-income
- This has consequences
- We need to think differently about test scores



Chapter 2:
Using—and Misusing—
Achievement Tests

The Challenge of Measuring Learning

- Learning does not have an inherent scale
- It is not like measuring height, weight, length and other physical properties
- Not to say we shouldn't try, but need to recognize the challenge

Three Key Dimensions of Test Scores

- Design
 - Norm-referenced vs. criterion-referenced
- **Reporting**
 - Raw scores, scale scores, ... (more below)
- Use
 - Low- vs. high-stakes
 - Formative vs. summative

Test Reporting

- Raw scores: # of items correct
- Scale scores: # correct, adjusted for item difficulty
- Developmental scale scores: particular kind of scaling that allows comparisons over time
 - Accomplished with linking items
- Normal-curve equivalents (NCE): Any of the above scores adjusted based on assumption that learning level has normal dist. (bell curve)

Interval Scale

- Ideally, would like to say that a one-point increase in scores means same thing no matter where we are on the test scale
- Called an “interval scale”
- Very difficult to accomplish
 - How can we “add up” learning on arithmetic and fractions?
 - How can we compare the amount of calculus learned to the amount of geometry?

“z-scores”

- Because scaling is imperfect, researchers often use z-scores
- In brief, this means assuming that the distribution of achievement doesn't change from grade to grade
 - z-scores are based on scale scores
- z-score of **zero** means that if all students started the year at the 50th percentile, they would remain there
 - Zero is average and does NOT mean no learning has occurred
- z-score **+0.10** means that a district/school/teacher moves students from the 50th to the 54th percentile
- This is reporting method planned for Missouri VA

Performance Standards

- Because of NCLB, we now focus on performance standards, esp. proficiency
- Amounts to setting a bar or “cut point” on the scale—very subjective
- Hard to judge learning based on performance standards because they throw out useful information
- Standardized tests are designed to be more precise near the cut points

Back to Snapshot Problem

- If you look only at snapshots, all of the above ways of reporting test scores are misleading for understanding educator performance
- But some reporting methods are better than others for VA, as I show next



Chapter 3: Measuring Student Growth

Why Growth?

- If the problem is accounting for what students bring with them to the classroom, then measure what they bring
- Annual student testing allows researchers to subtract prior scores from current ones—growth
- Growth can be calculated for different test score reporting methods
- Ideal: Growth of individual students based on scale scores or NCEs (or z -scores)

Illustration #1: Teachers w/ Same Growth

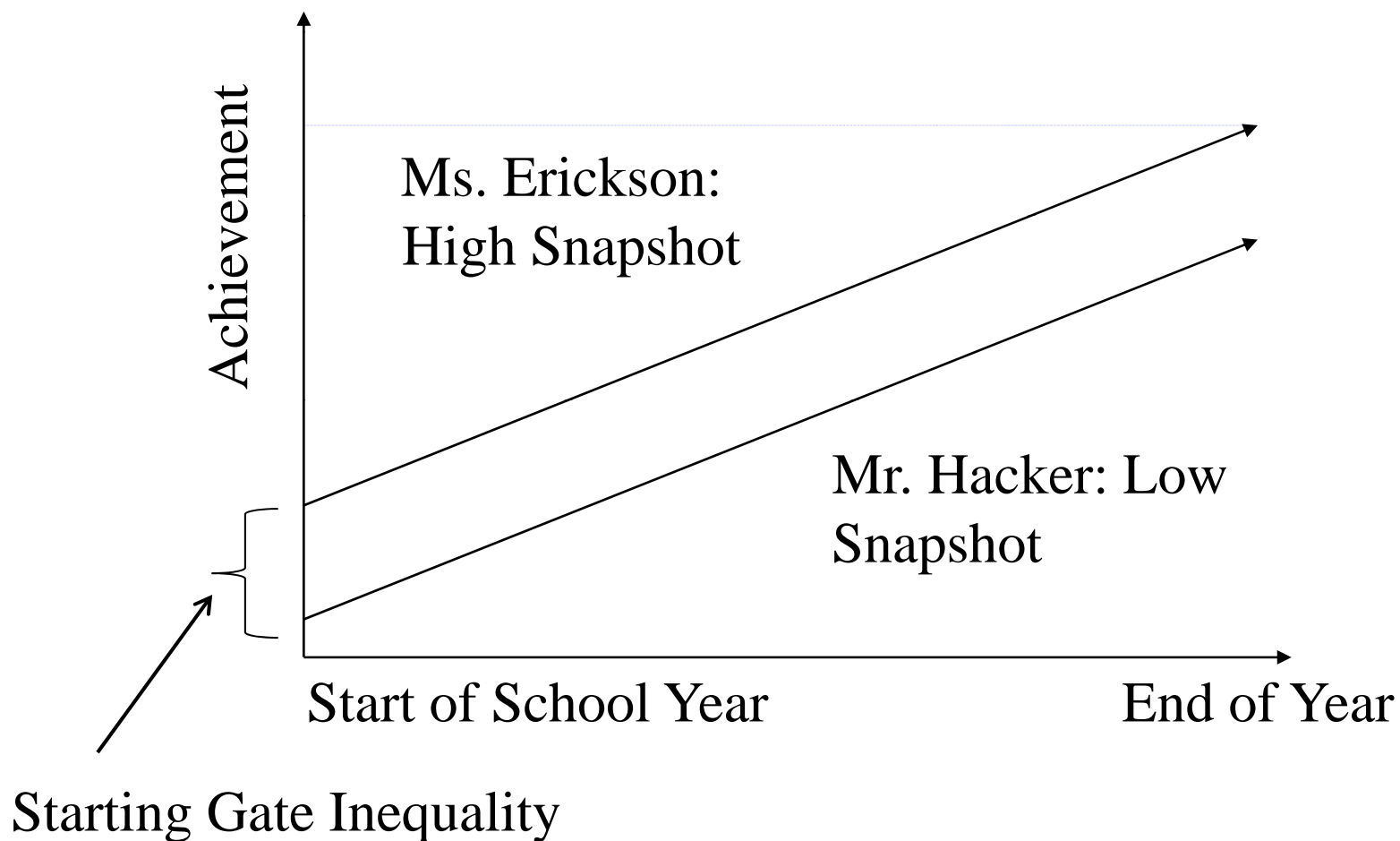
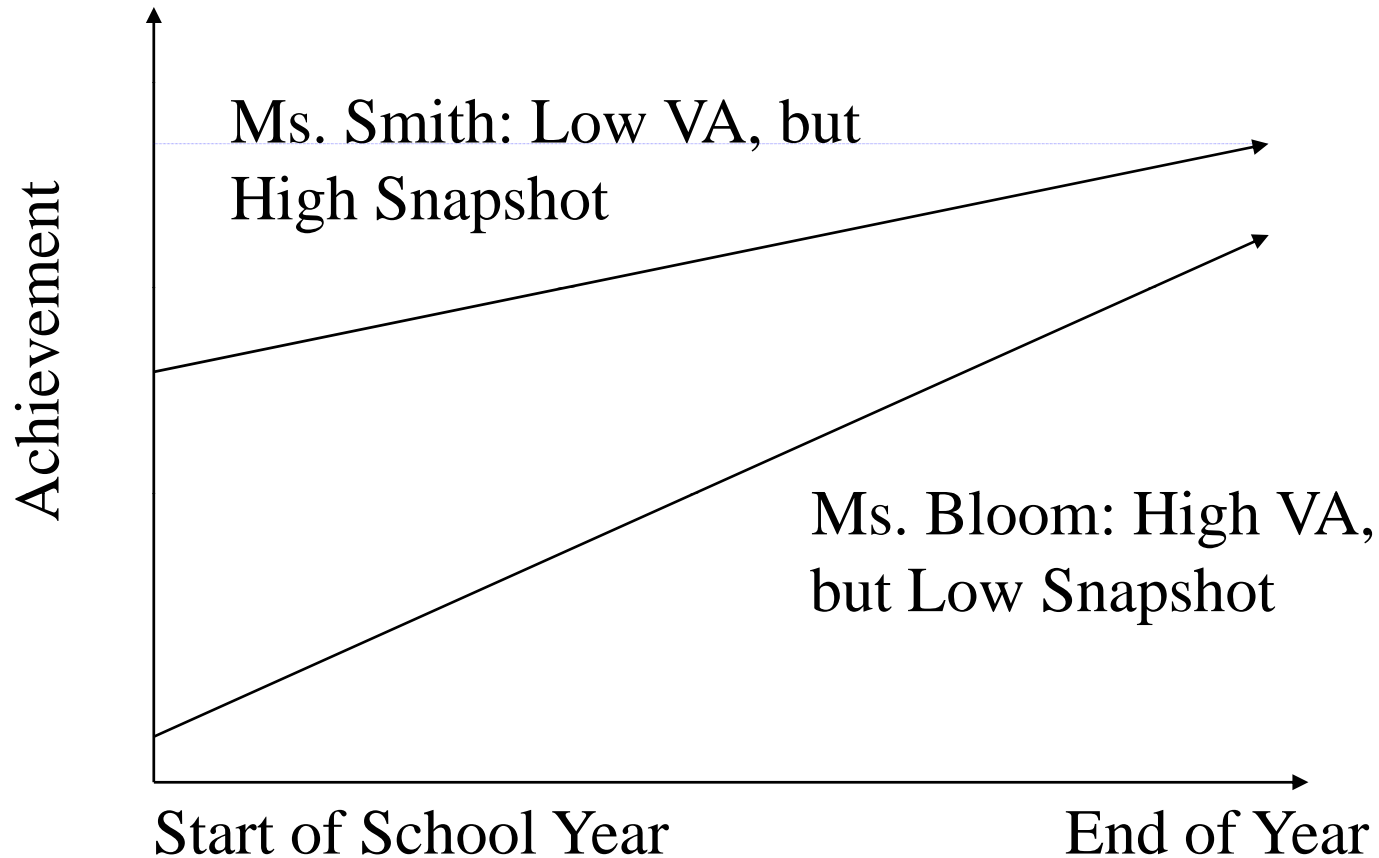
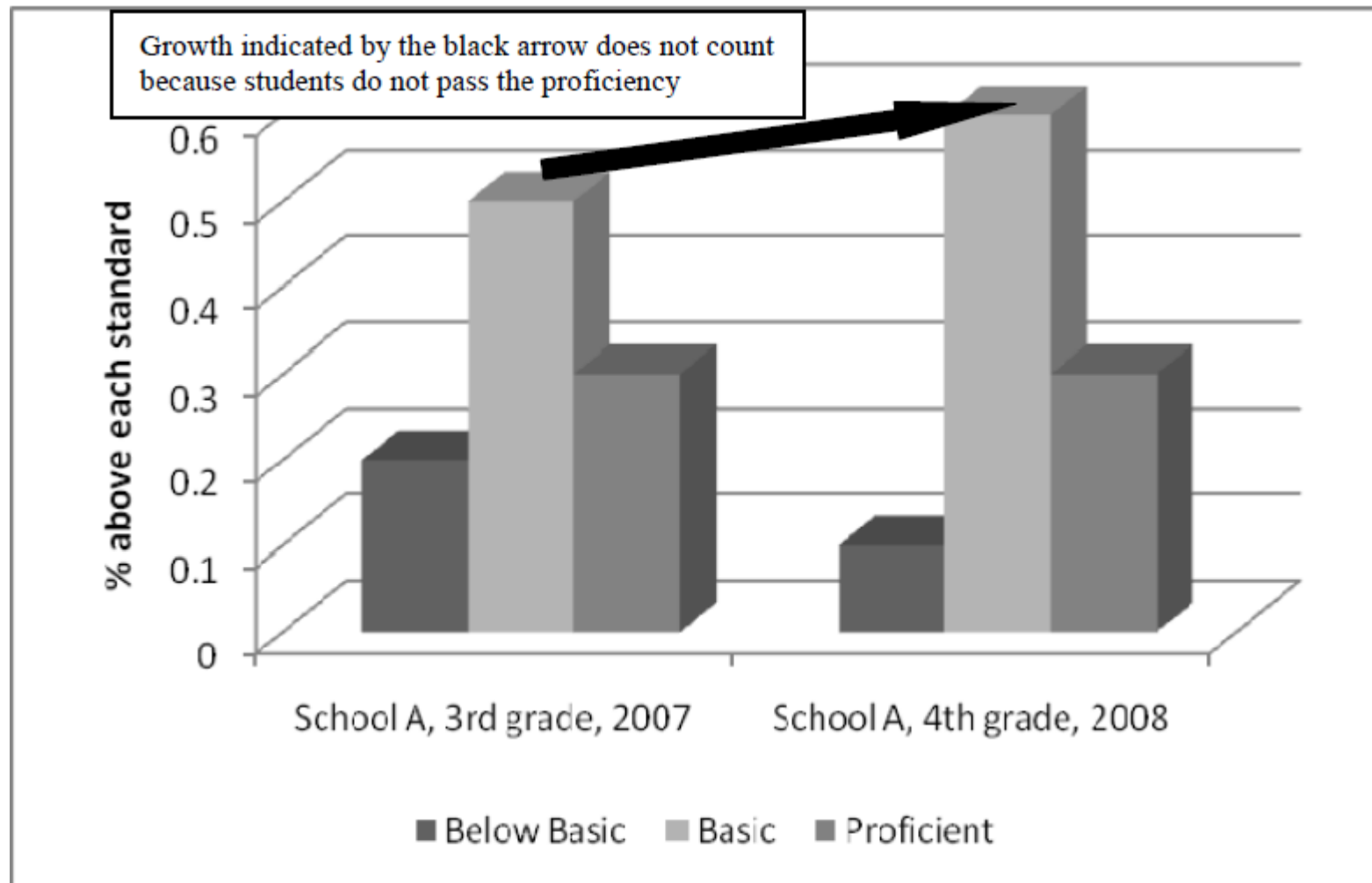


Illustration #2: Teachers w/ Different Growth

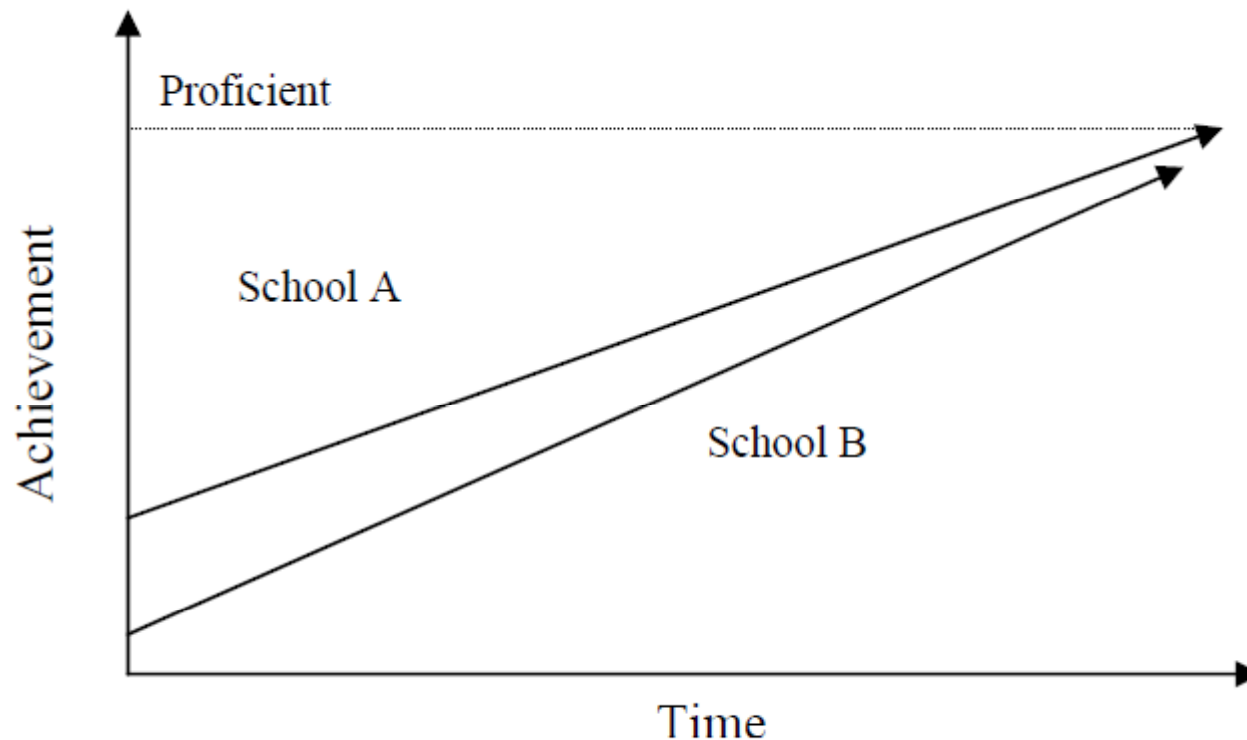


“Growth” with Proficiency Standards



More Things NOT Growth

- NCLB “Adequate Yearly Progress”
- Federal growth pilot (Missouri a pilot state)



Summary of Diff. Forms of Growth

<i>Type of Growth</i>	<i>Test Reporting Method</i>		
	<i>Scale Scores</i>	<i>Normal Curve Equivalents</i>	<i>Performance Standards</i>
<i>Student Growth</i>	Best Approach for Measuring School Performance if scaling procedure is trustworthy	Best Approach for Measuring School Performance if scaling procedure is questionable	* Wastes information
<i>Cohort-to-Cohort Growth</i>	* Captures irrelevant changes in student cohorts	* Captures irrelevant changes in student cohorts	* Captures irrelevant changes in student cohorts * Wastes information
<i>Growth-to- Proficiency</i>	* Holds school accountable for factors outside their control	* Holds school accountable for factors outside their control	* Holds school accountable for factors outside their control * Wastes information

Student Growth Percentiles (SGP)

- SGPs report scale score growth in terms of percentiles
 - Confusing point: “Growth in percentiles” not same as “Percentile of growth” (SGP)

Not Just a “Measurement” Problem

- Shifting from snapshots to growth doesn't just change the measure, it changes the way we look at education
- Education should be about taking students where they are and moving them forward as much as possible
- This implies focus on growth measures



Chapter 4: Creating Value-Added Measures

From Growth to VA

- Growth is first step toward value-added
- Limits of simple growth:
 - Unequal school resources
 - Prior achievement may not be enough to account for student differences
- Possible solution: Compare similar schools
 - Put into buckets w/ similar resources and students
 - Apples to apples (within buckets)
- Teachers whose students make greater than average growth in bucket are “high value-added”

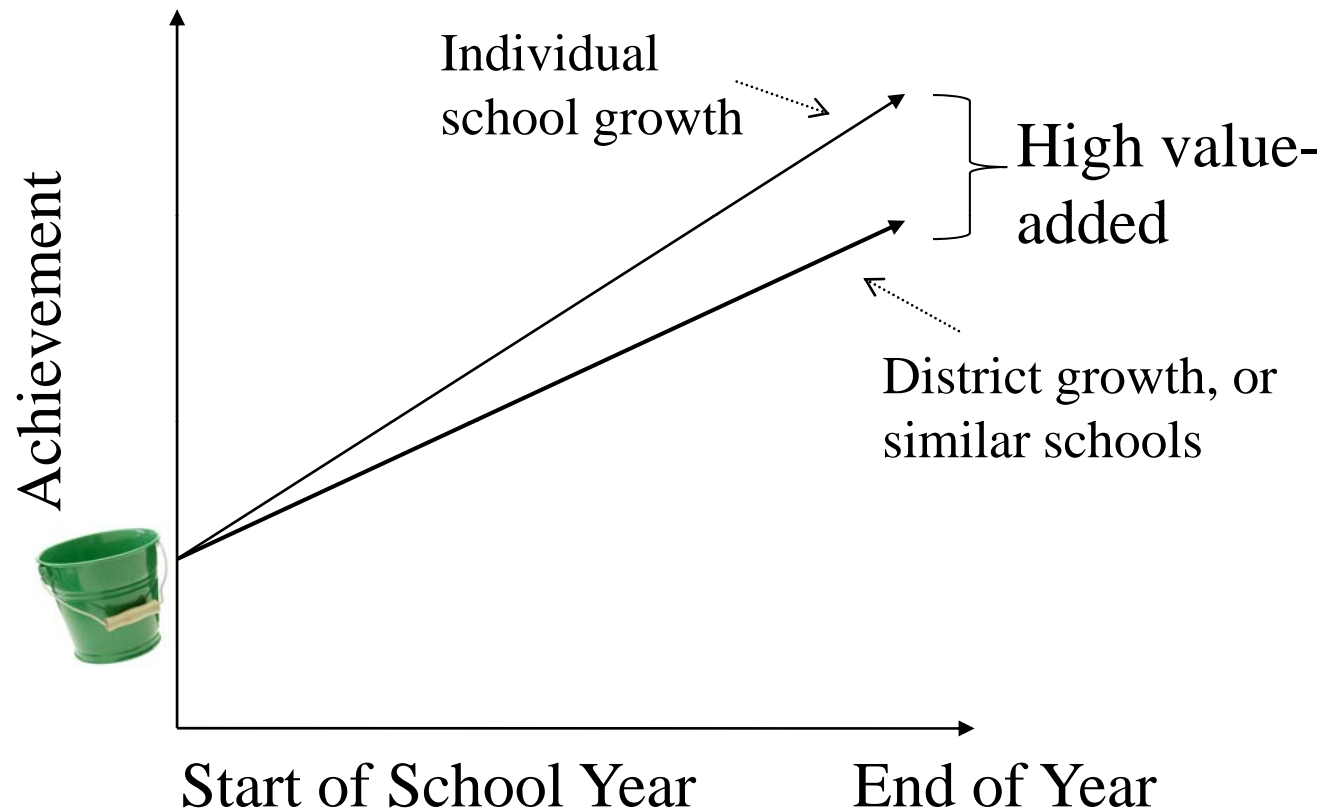


High
Prior
Achieve.

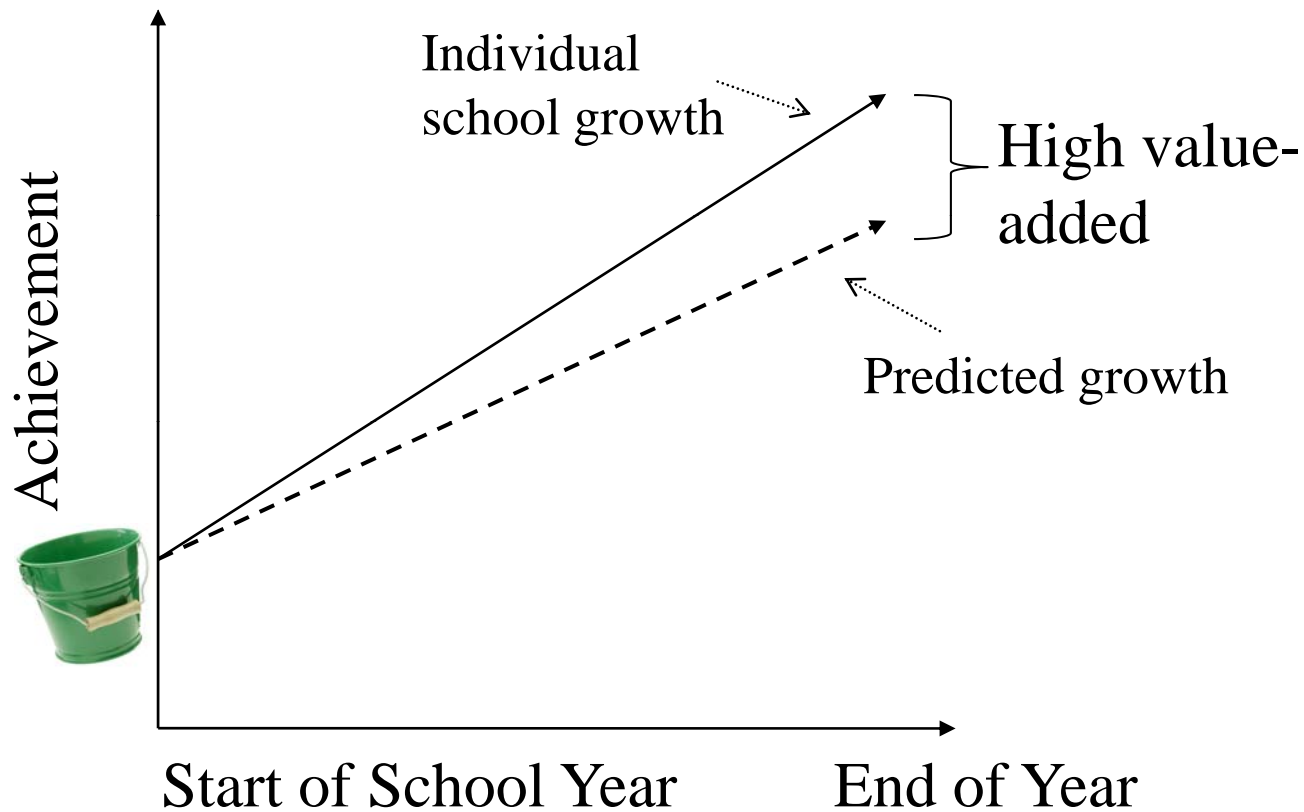


Low
Prior
Achieve.

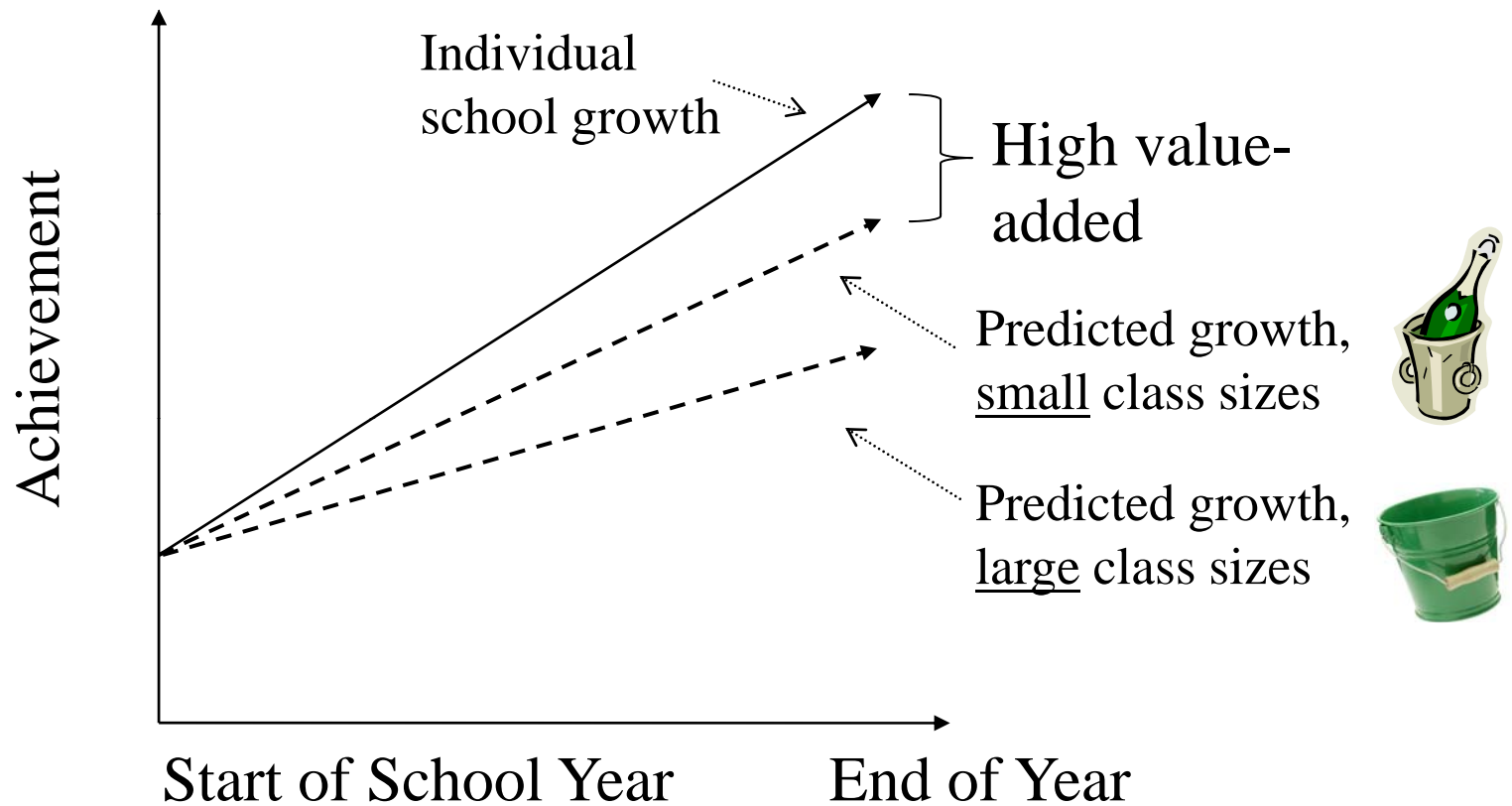
VA Illustration #1: Comparison Appr.



VA Illustration #2: Prediction Approach



VA with “Controls”



Summary of Controls

<i>Uncontrollable</i>		<i>Partly Controllable</i>
<i>Measured (Maybe account for)</i>	<i>Unmeasured (May want to account for but cannot)</i>	<i>Measured (Shouldn't account for)</i>
<i>School Resources</i> Class size Funding Staff positions <i>Student Factors</i> Prior test scores Race/income (?) Mobility Variation in snapshot achievement	District leadership District funding District policies Collaboration among schools	Teacher credentials Teacher experience Student absences Program participation (e.g., special ed, gifted)

Value-Added Measures are Relative

- VA allows us to make comparisons among schools and teachers (it's relative), not draw absolute conclusions about performance
- On the one hand, this means that some teachers and schools will have low value-added no matter what they do
- On the other hand, we would never want to say that when a teacher or school gets to a particular standard, that they are “good enough”
 - Relative measures facilitate continuous improvement

VA vs. Student Growth Percentiles (SGP)

- SGPs often described as “descriptive” and VA criticized for making “attributions” and “blaming,” but this is a false distinction
 - It is logically impossible to respond to “descriptive” information without making an attribution
 - Policymakers and others may place blame, but this has nothing to do with whether we use VA or SGPs



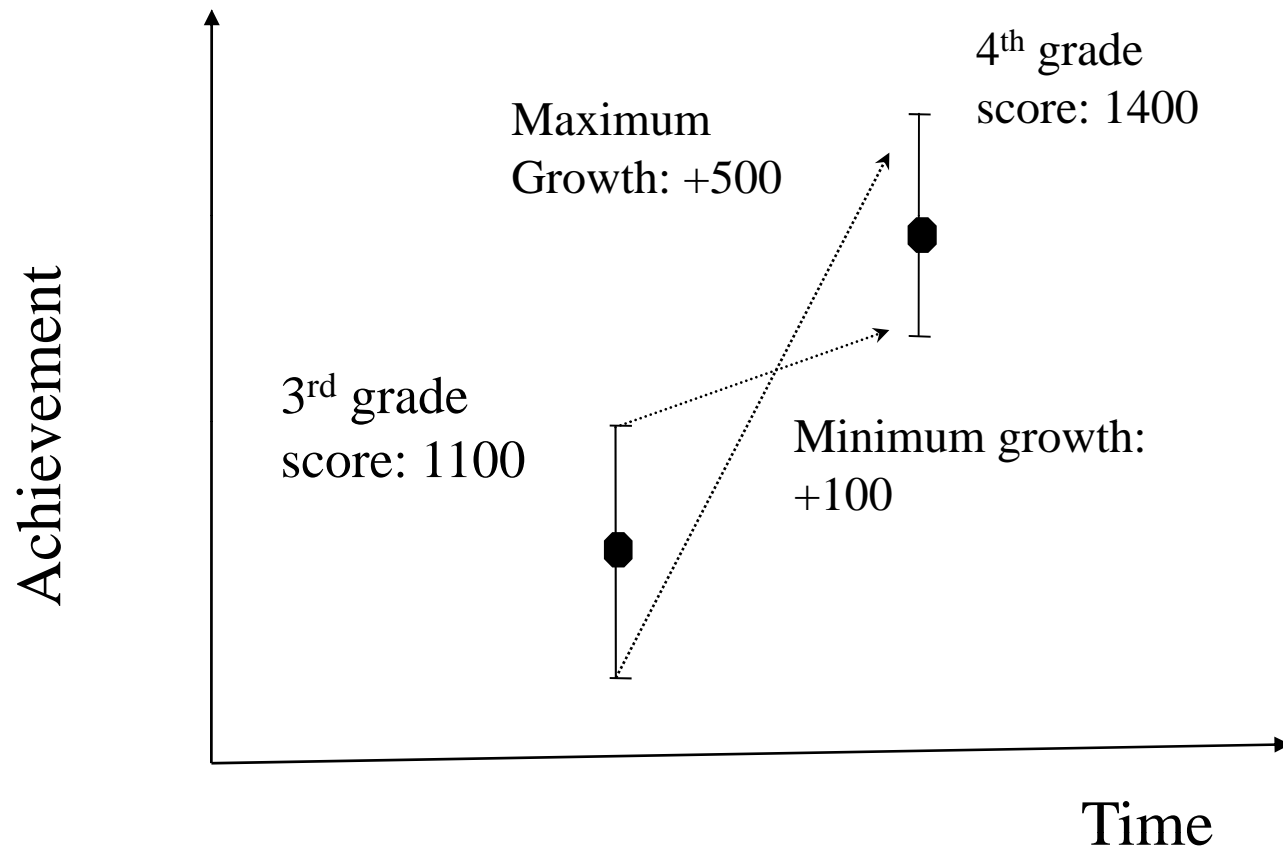
Now, for the problems . . .

Chapter 5:
Understanding Statistical Errors in
Value-Added

Two Basic Types of Errors

- Systematic Error: More likely to occur with a particular school or teacher
 - Snapshots are a case in point: they systematically disadvantage low-snapshot schools
- Random Error: Equally likely for everyone
 - Example: A coin toss
 - Two sources:
 - Measurement error (from the student test scores)
 - Sampling error (more students, less sampling error)
 - Random error is worse with growth measures

Random Error Worse with Growth Measures



Relationships Between Errors

- (1) Random error in an individual student's score:
 - (a) increases random error in educator value-added performance measures; but
 - (b) does not increase systematic error in educator value-added measures.

Relationships Between Errors (cont.)

(2) Systematic error in an individual student's score does not introduce systematic error into value-added performance measures if:

(a) systematic error is constant across years for individual students and therefore cancels out for those students (test anxiety); or

(b) systematic error in student scores is evenly distributed across classrooms and schools (stereotype threat with girls and their math and science scores)

Two Parting Thoughts on Errors

- Errors in statistics do not necessarily lead to errors in decisions, though more likely
 - Depends on policies about use of measures
- “We made too many wrong mistakes”

-- Yogi Berra

What is Missouri doing with VA? Part I

- School- and district-level VA this fall (pilot)
 - Grades 4-8
 - Based on state MAP test
- Teacher-level value-added in 2012
- No decisions made about including student demographics as controls (to identify similar schools)
- VA measures will be reported as z -scores along with confidence intervals

Summary and Preview

- The goal today has been to explain the basic logic behind value-added, as well as what can go wrong (potential errors)
- On Friday, I will focus on the later chapters of the book: evidence about value-added measures and recommendations about how to create and use the measures
- Now, it's time for questions . . .