Learning From Research on Test Based Accountability

Daniel Koretz
Harvard Graduate school of Education

World Bank Symposium:
Assessment for Global Learning

November 7, 3013

Problem statement

- Growing interest worldwide in using tests for evaluation and test-based accountability (TBA)
  - Now increasing pressure to set targets in post-2015 context
- Substantial experience with “high-stakes” testing in US
  - Many programs since early 1970s
  - Research evaluating impact since late 1980s
- Research (mostly in US) shows serious problems
  - Still little research elsewhere
- Need to build systems that minimize these problems

"High stakes:" pressure to raise scores

- Pressure can come from many sources:
  - Tangible consequences for students
  - Tangible consequences for individual teachers
  - Tangible consequences for schools as organizations
  - Publicity and public pressure
- Research shows that effects can arise with any of these
- Limited evidence that severity of stakes—overall or for specific groups—influences severity of responses

What we know about high-stakes testing

- Effects on educational practice are mixed
  - Some improvements
  - Many undesirable effects—bad test preparation, other "gaming"
- Scores can become severely inflated (increase much more than actual learning)
  - Overall improvement is exaggerated—often severely
  - Relative effectiveness is misestimated
    - Effects on equity misestimated
    - Teachers and schools ranked incorrectly

What we don't know

- What is the net effect on student achievement?
  - Weak research designs, weaker data
  - Some evidence of inconsistent, modest effects
  - Effects are likely to vary across contexts
- Which types of test-based accountability systems are best?
  - Which programs maximize real improvements
  - Which programs minimize gaming, bad test preparation, & score inflation
- Reason: grossly inadequate research and evaluation

“Campbell’s Law” (1975)

“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.”

Donald T. Campbell, (1975). “Assessing the impact of planned social change.” In G. M. Lyons (Ed.), *Social Research And Public Policies : The Dartmouth/OECD Conference*.

Examples of Campbell's Law

- Airline on time statistics
- West Virginia postal delivery times
- Cardiology "report cards" in New York

For many more examples, see:

https://my.vanderbilt.edu/performanceincentives/files/2012/10/200804_Rothstein_HoldingAccount.pdf

The sampling principle of testing: analogy of a political poll (Colombia)

- On 3 June 2010, a poll of 2,000 people poll by Centro Nacional de Consultorí predicted 61.6% for Santos, 29.8% for Mockus
- Actual vote: 69.1% for Santos, 27.5% for Mockus
- Would you have cared how those particular 2,000 people actually voted?
- Why is information from those 2,000 people valuable?

What are the consequences of incomplete sampling?

- All cases:
  - Systematically incomplete evaluation of education
- Low pressure: modest effects on scores
  - Measurement error (uncertainty): fluctuations in scores
  - Differences in results among tests: usually modest, but not always (TIMSS vs. PISA, US NAEP vs. TIMMS)
- High pressure (accountability): very large effects
  - Incentives to focus on the tested sample, not the domain
  - Narrowed instruction, bad test preparation
  - Score inflation

Why inflation occurs

- Tests show predictable emphases, omissions, and forms of presentation over time.

- Test prep can capitalize on these recurrences:
  - Reallocation: aligning instruction to focus on emphasized content, at the cost of other content relevant to the inference
  - Coaching: focusing on presentation, rubrics, and incidental test content
  - Cheating

Algebra 1

7.1			
7.2	2003S #17 (o)		
7.3			
7.4	2003S #38 (m)	2002F #37 (m)	2000S #36 (m)
7.5			
7.6			
7.7			

Source: Quincy MA High School Math Dept.

How similar are tested representations?

2009

9 Which tool would be **most appropriate** for Natasha to use when finding the mass of a watermelon?

A scale

B inch ruler

C meter stick

D measuring cup

2007

27 Which tool is **most** appropriate for measuring the mass of a serving of cheese?

A ruler

B thermometer

C measuring cup

D weighing scale

How similar are tested representations?

NY 7N7: Compare numbers written in scientific notation.

2 The table below shows the number of computers a company sold in four different years.

COMPUTERS SOLD

Year	Computers Sold
2002	3.2×10^5
2003	8.4×10^3
2004	5.9×10^5
2005	1.2×10^4

In what year did the company sell the **most** computers?

A 2002

B 2003

C 2004

D 2005

18 Connor is researching four types of memory modules for his computer. The data are shown in the table below.

Module	Amount of Memory (in bytes)
W	3.64×10^8
X	1.28×10^9
Y	2.56×10^9
Z	5.12×10^8

Connor wants to buy the module with the most memory. Which module should he buy?

A Module W

B Module X

C Module Y

D Module Z

An example of coaching (cheating?)

"The question on the review sheet for…[the] exam…reads in part:

'The average amount that each band member must raise is a function of the number of band members, b, with the rule f(b)=12000/b.'

The question on the actual test reads in part:

'The average amount each cheerleader must pay is a function of the number of cheerleaders, n, with the rule f(n)=420/n'."

Strauss, V., *The Washington Post*, July 10, 2001, p. A09

Coaching: based on an incidental characteristic of test items

Whenever you have a right triangle—a triangle with a 90-degree angle—you can use the Pythagorean theorem…. the sum of the squares of the legs of the triangle (the sides next to the right angle) will equal the square of the hypotenuse (the side opposite the right angle)….

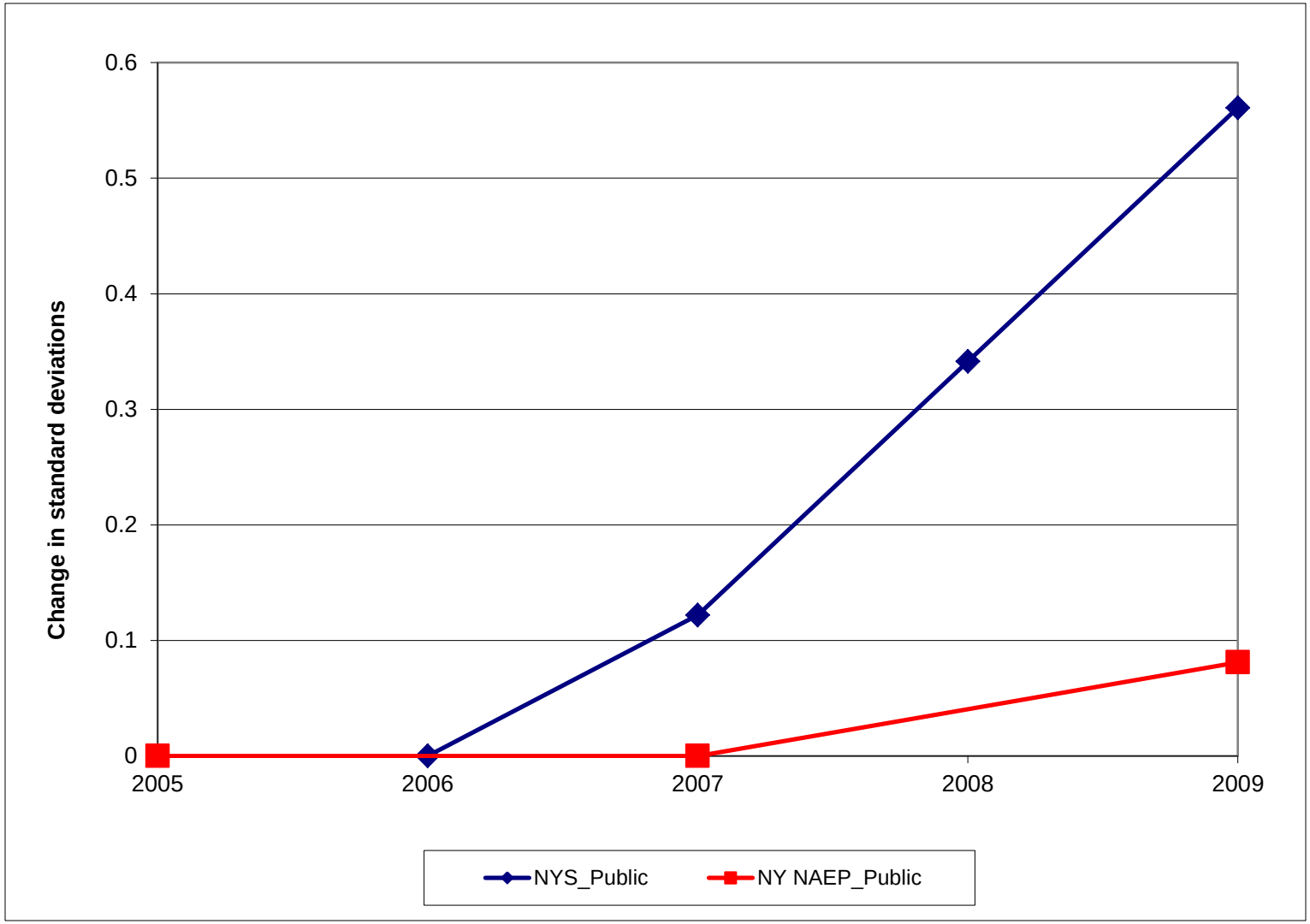
Two of the most common ratios that fit the Pythagorean theorem are 3:4:5 and 5:12:13. Since these are ratios, any multiples of these numbers will also work, such as 6:8:10, and 30:40:50.

Princeton Review, *Cracking The MCAS Grade 10 Mathematics*

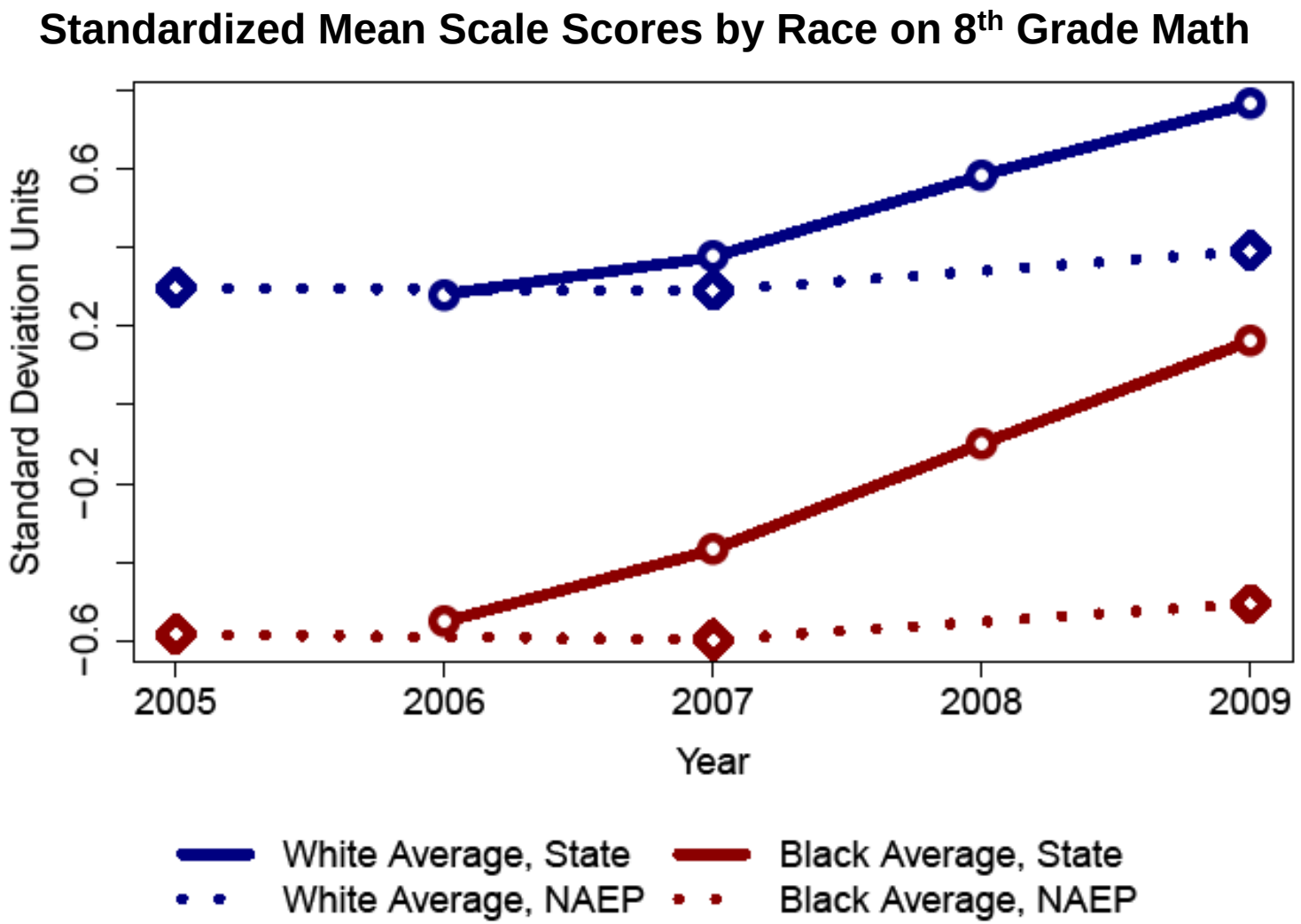
Logic of studies of score inflation

- Scores are meaningful **only** if they generalize to the domain
  - A poll is useful only if its results generalize to the entire electorate
- If gains generalize to the domain, they must generalize to other tests of the same domain
  - If a poll is accurate, other polls will show similar results

Grade 8 math score trends in New York State



Trends by Race on New York State vs. NAEP



Can inflation affect international large-scale assessments?

- Short answer: yes
- Assessments without scores for individuals can be designed to give broader coverage (matrix sampling)
  - Makes inappropriate test preparation modestly harder
- But primary protection is that in many countries, there is no pressure to prepare specifically for them
- Where there *is* pressure to prepare specifically for ILSAs *or tests modeled after them*, inflation can arise
- To my knowledge, no independent studies yet of ILSAs examining:
  - Degree of predictability
  - Behavioral responses to testing
  - Score inflation

How to do better?

- Be cautious: take advantage of problems shown by the U.S. experience
- Keep a broader focus: not just test scores
- Avoid unrealistic performance targets
- Design tests appropriately for use in accountability systems
- Monitor and evaluate the *testing and evaluation* system routinely, and be prepared to modify them

Need to experiment with new system designs

- Need to find ways to make other goals *count*
  - Including higher-order skills that are hard to assess with an externally imposed test
- Need to explore the use of multiple measures
  - Additional objective measures
  - Subjective measures
- Need to monitor for gaming, consider “dynamic” accountability

Need to experiment with new test designs

To better estimate real gains and improve incentives

- Maximize breadth of coverage
- Minimize unnecessary repetition of:
  - Content
  - Styles of presentation
  - Task demands and scoring
- Build in “audit” testing
  - In sample-based testing program
  - With embedded items (“self-monitoring assessments”)

Need monitoring of evaluation systems

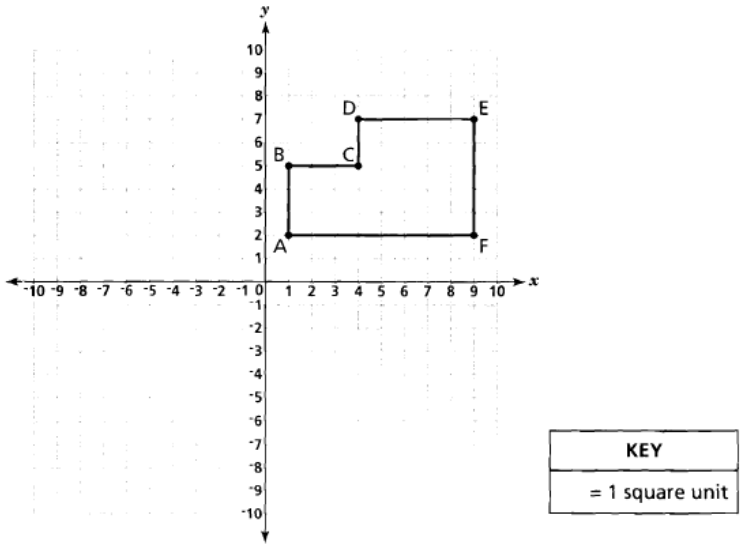
- Need monitoring of:
  - Behavioral responses by educators
  - Other forms of gaming
  - Score inflation
- Need investigation of variations in effects, for example:
  - Variations across types of schools
  - Variations across types of students

Supplementary slides

How similar are tested representations?

6G11: Calculate the area of basic polygons drawn on a coordinate plane (rectangles and shapes composed of rectangles having sides of integer length)

17 Figure ABCDEF is plotted on the coordinate plane below.



KEY
= 1 square unit

What is the area, in square units, of the figure?

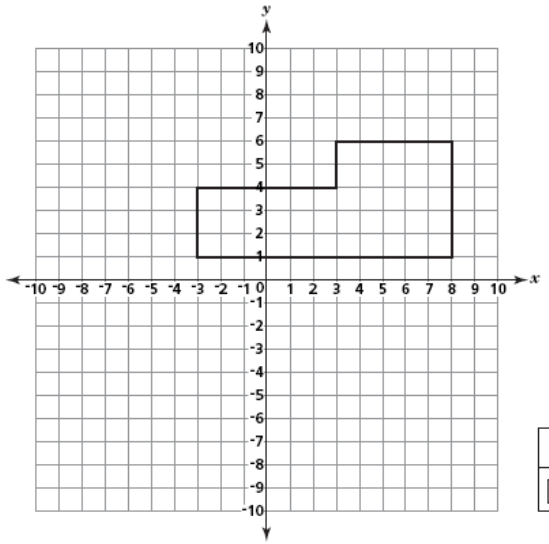
A 40

B 34

C 26

D 25

4 A polygon is plotted on the coordinate plane below.



KEY
= 1 square unit

What is the area, in square units, of the polygon?

A 25

B 32

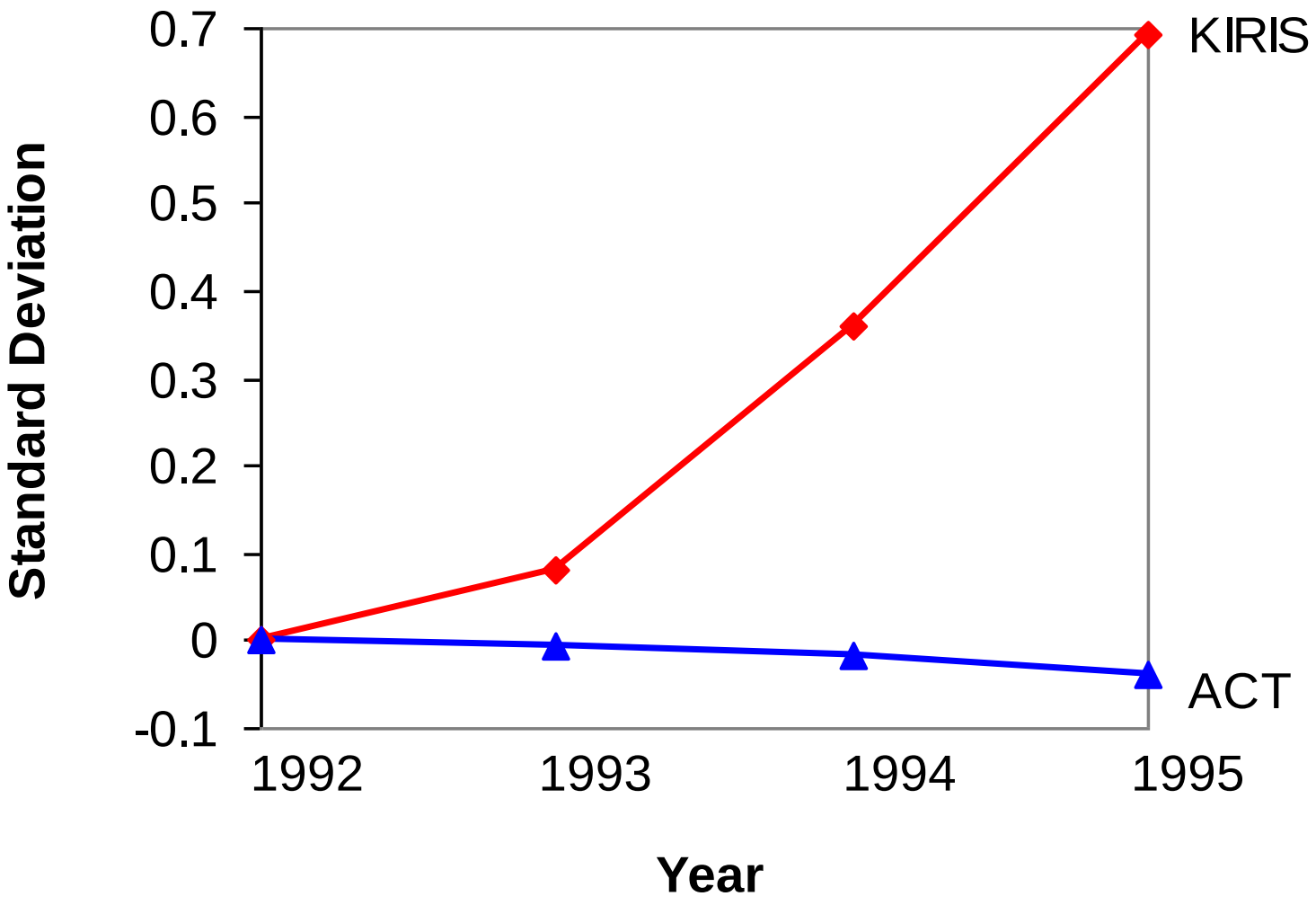
C 43

D 55

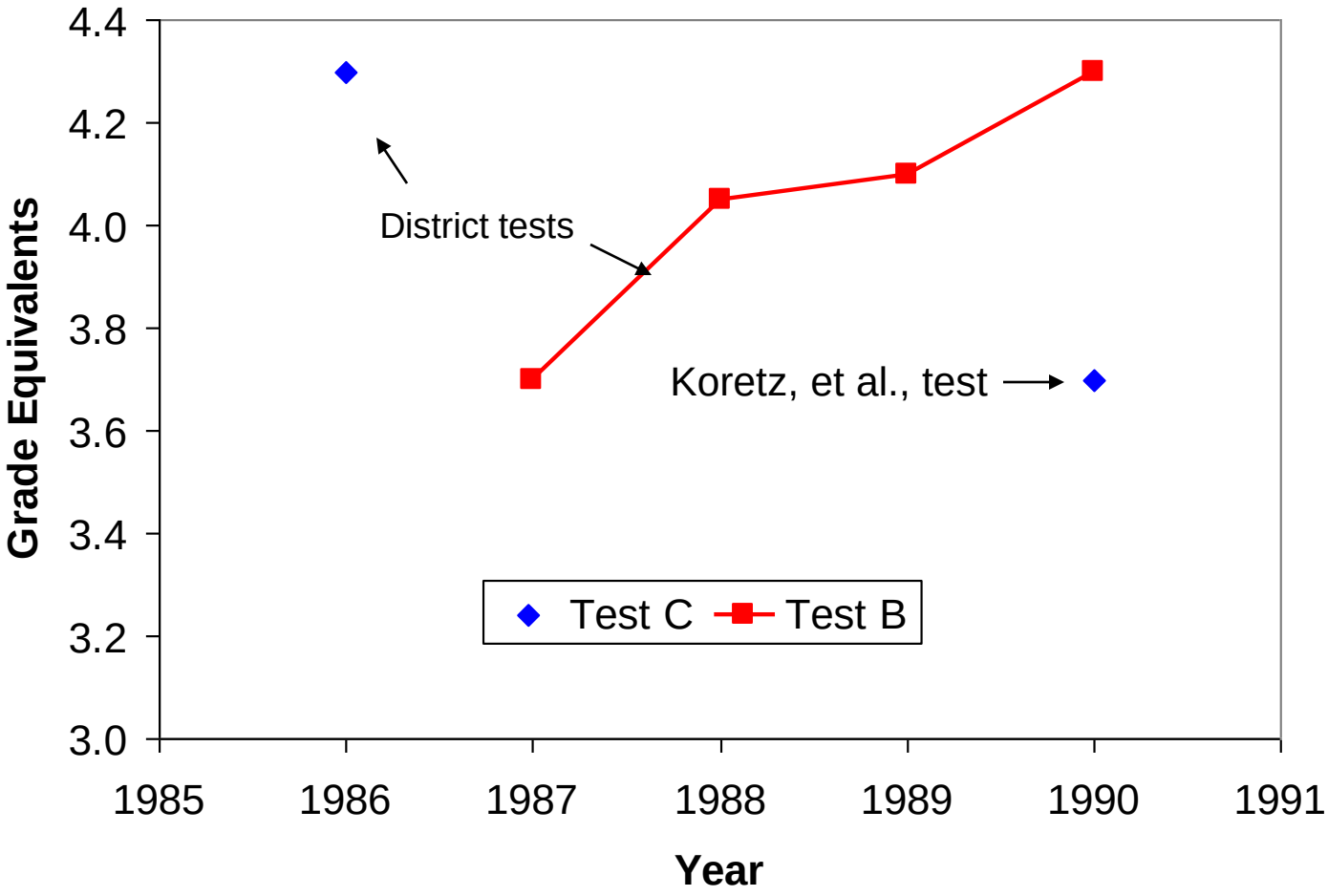
Reading change, grade 4 KIRIS and NAEP, 1992-1994

	KIRIS	NAEP
Gain in scale scores	18.8	-1
Standardized Gain	0.76	-0.03

Math trends, KIRIS and ACT



Performance on coached and uncoached tests



SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)

Good versus bad preparation for a test

- Good: gives students knowledge and skills that they can apply elsewhere
  - In later education
  - In later employment
  - Therefore, on other tests
- Bad: generates test-specific gains that do not generalize beyond that test