Using Large-Scale Assessments to Improve Education: Lessons from the U.S. Experience

Daniel Koretz
Harvard Graduate school of Education

National Institute for the Evaluation of Education/
World Bank Symposium:

Towards a New Generation of Standardized Student Assessments

June 23, 2014
Mexico City

Problem statement

- Growing interest worldwide in using large-scale assessments to improve education
- Substantial experience with “high-stakes” testing in U.S. (testing with accountability for scores)
- Evaluations of U.S. programs provide guidance for the development of better systems elsewhere

Key aspects of the U.S. system

- Policies using test-based accountability to improve instruction have been in place for decades
  - State policies until 2001
  - Federal *No Child Left Behind* law since 2002
- Different forms of assessments tried, for example:
  - Performance assessments
  - "Standards-based" assessments
- Different types of accountability, for example:
  - Using current scores (status) versus "value-added"
  - Sanctions and rewards for schools, for individual teachers, or for students
- Performance targets were arbitrary, often very high, and often uniform

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 estimated incorrectly
    - Teachers, schools, and systems ranked incorrectly
    - Can create an illusion of greater equity

What we don't know

- What is the net effect on student achievement?
  - Weak research designs, weaker data
  - Some evidence of inconsistent, modest effects in elementary math, none in reading
  - 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

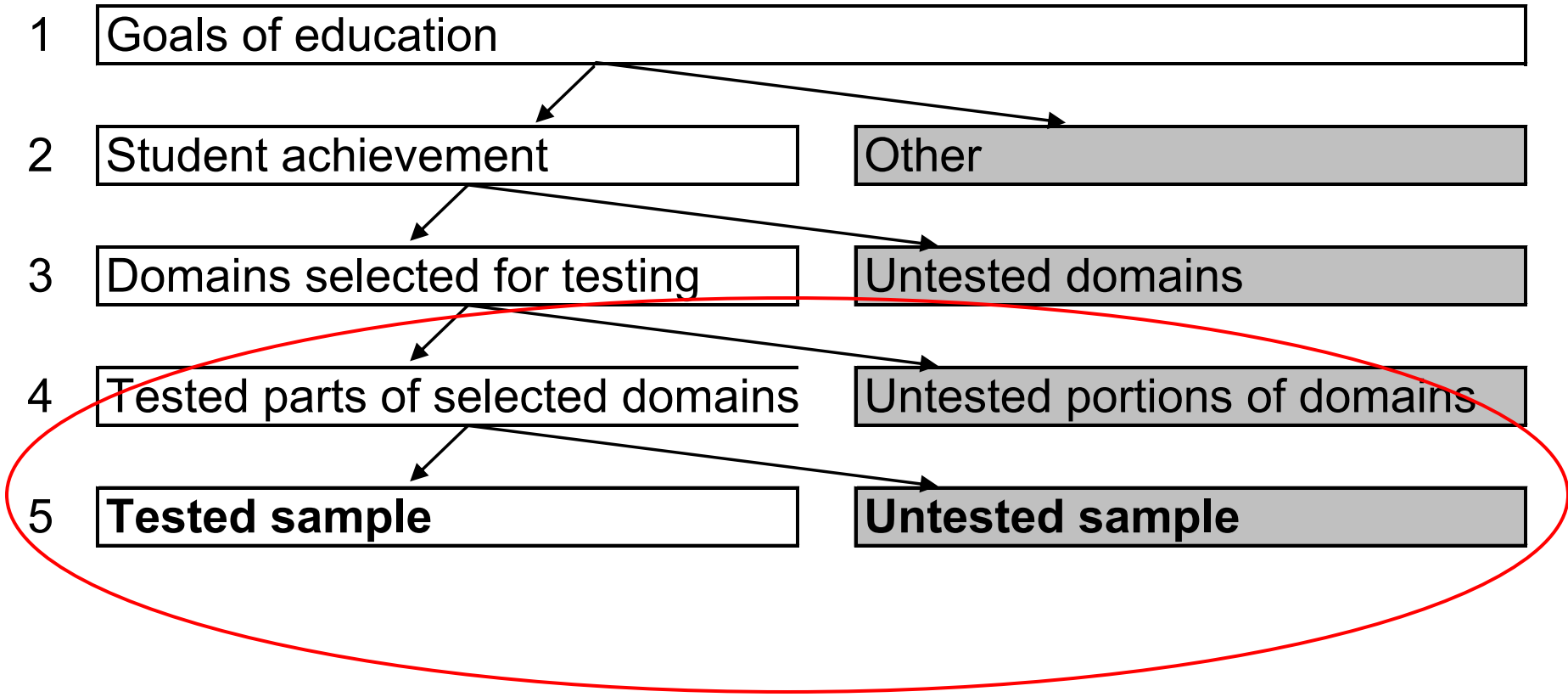
Topics

- The “sampling principle” of testing
- Score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new testing and evaluation programs

The sampling principle of testing: analogy of a political poll

- June 2012 poll by Berumen y associados predicted 33.4% for Peña Nieto, 27.3% for Lopez Obrádor
- Actual vote: 38.2% for Peña Nieto, 31.6% for Lopez Obrádor
- Would you have cared how the few *specific* people polled by Berumen voted?
- Why is information from these few polled people valuable?

Sampling to obtain a test



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, for example, TIMSS vs. PISA
- High pressure (accountability): very large effects
  - Incentives to focus on the tested sample, not the domain
  - Narrowed instruction, bad test preparation
  - Score inflation

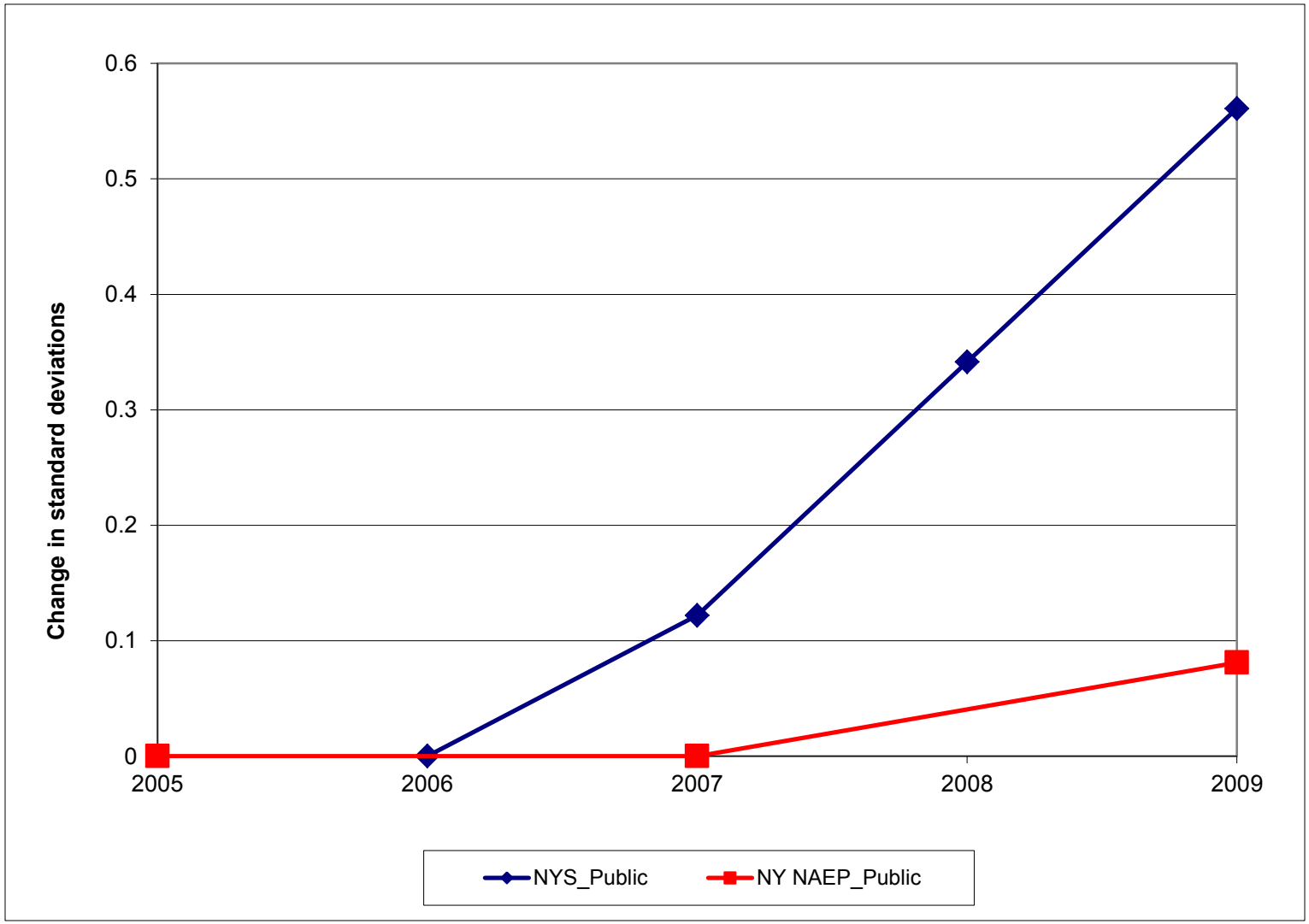
Topics

- The “sampling principle” of testing
- Score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new testing and evaluation programs

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
  - Gains on a high-stakes test should generalize to a lower-stakes "audit" test
  - If a poll is accurate, other good polls will show similar results

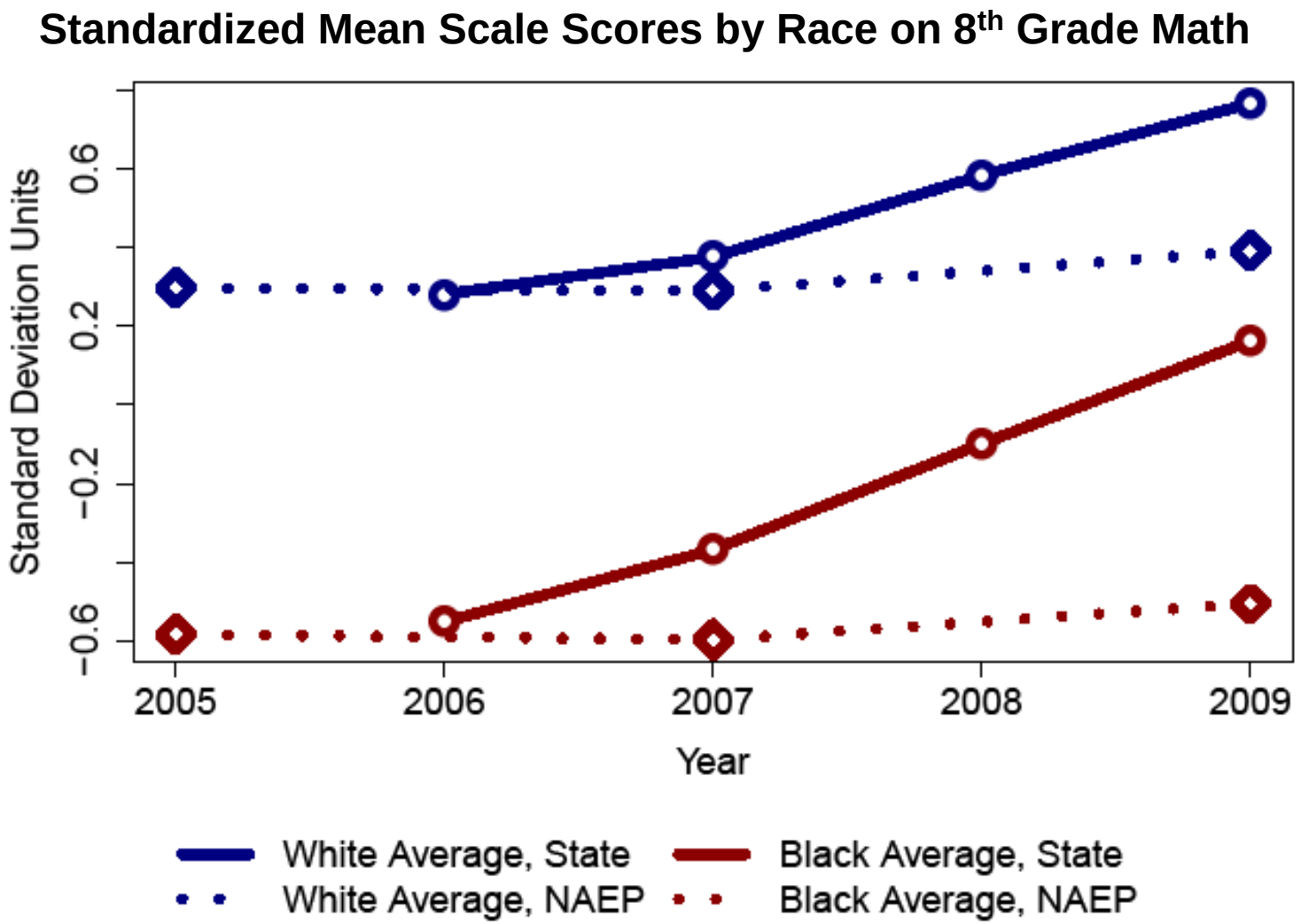
Grade 8 math score trends in New York State



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

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

Trends by Race on New York State vs. NAEP



Topics

- The “sampling principle” of testing
- Score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new testing and evaluation programs

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 score inflation: test-specific gains that do not generalize beyond that test

Why inflation occurs

- Tests show predictable emphases, omissions, and forms of presentation over time.
  - Some intentional, for technical reasons
  - Some accidental or to save time and money

- Test preparation can capitalize on these patterns:
  - Reallocation: focusing instruction 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 Massachusetts High School Math Department

Coaching

- Focusing preparation on substantively unimportant details of the test
  - Minor, unimportant details of content
  - Details of the presentation of material
- Includes test-taking tricks (e.g., process of elimination, plug-in)
- Can inflate scores or simply waste time

c. meter stick
d. measuring cup

From 2008 (Standard 7M9)

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

a. ruler
b. thermometer
c. measuring cup
d. weighing scale

2009 item, New York grade 7 math test

From 2009 (Standard 7M9)

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

a. scale
b. inch ruler
c. meter stick
d. measuring cup

From 2008 (Standard 7M9)

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

a. ruler

Item from G8 MCAS

Eva has four sets of straws. The measurements of the straws are given below. Which set of straws could not be used to form a triangle?

A. Set 1: 4 cm, 4 cm, 7 cm
B. Set 2: 2 cm, 3 cm, 8 cm
C. Set 3: 3 cm, 4 cm, 5 cm
D. Set 4: 5 cm, 12 cm, 13 cm

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*

Topics

- The “sampling principle” of testing
- Score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new testing and evaluation programs

What we have learned from the U.S. experience

- Test-based accountability has not worked very well:
  - Tests omit many important outcomes
  - High-stakes testing generates mixed effects on practice
  - High-stakes testing produces inflated score gains
- Score inflation undermines evaluation in two ways:
  - Overall improvement is exaggerated
  - Relative effectiveness (for example, of schools) is estimated incorrectly

Recommendation 1: couple evaluation and accountability with training and support

- Many teachers need help, not just incentives, to improve instruction
- Accountability for performance should be accompanied by training and other supports for improving instruction

Recommendation 2: make the evaluation and accountability system broader

- Do not rely only or excessively on standardized tests
- Evaluate other outcomes
- Evaluate *practices* as well as *outcomes*
- May need to use *subjective* as well as *objective* measures

Recommendation 3: Use summative tests appropriately

- Set *realistic* performance targets that teachers can reach by appropriate methods
  - Creates less incentive to use bad test preparation
- Report in terms of scale scores, not performance standards
  - Performance standards alone create bad incentives, misleading analysis, and misunderstanding

Recommendation 4: design tests for accountability uses

- Avoid excessive narrowing in summative tests
- Avoid excessive predictability in summative tests:
  - Content
  - Format/presentation
  - Task demands (for example, rubrics)
- Design formative tests *differently*
  - To serve formative purpose
  - To avoid undesirable test preparation

Recommendation 5: evaluate the evaluation and accountability system

- No proven system for doing all of this
- *All* accountability systems cause some undesirable responses
- Therefore, increases in scores are not enough to indicate success
- It is essential to monitor the effects of the accountability system

Evaluating the evaluation system

- Need monitoring of:
  - 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

"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."

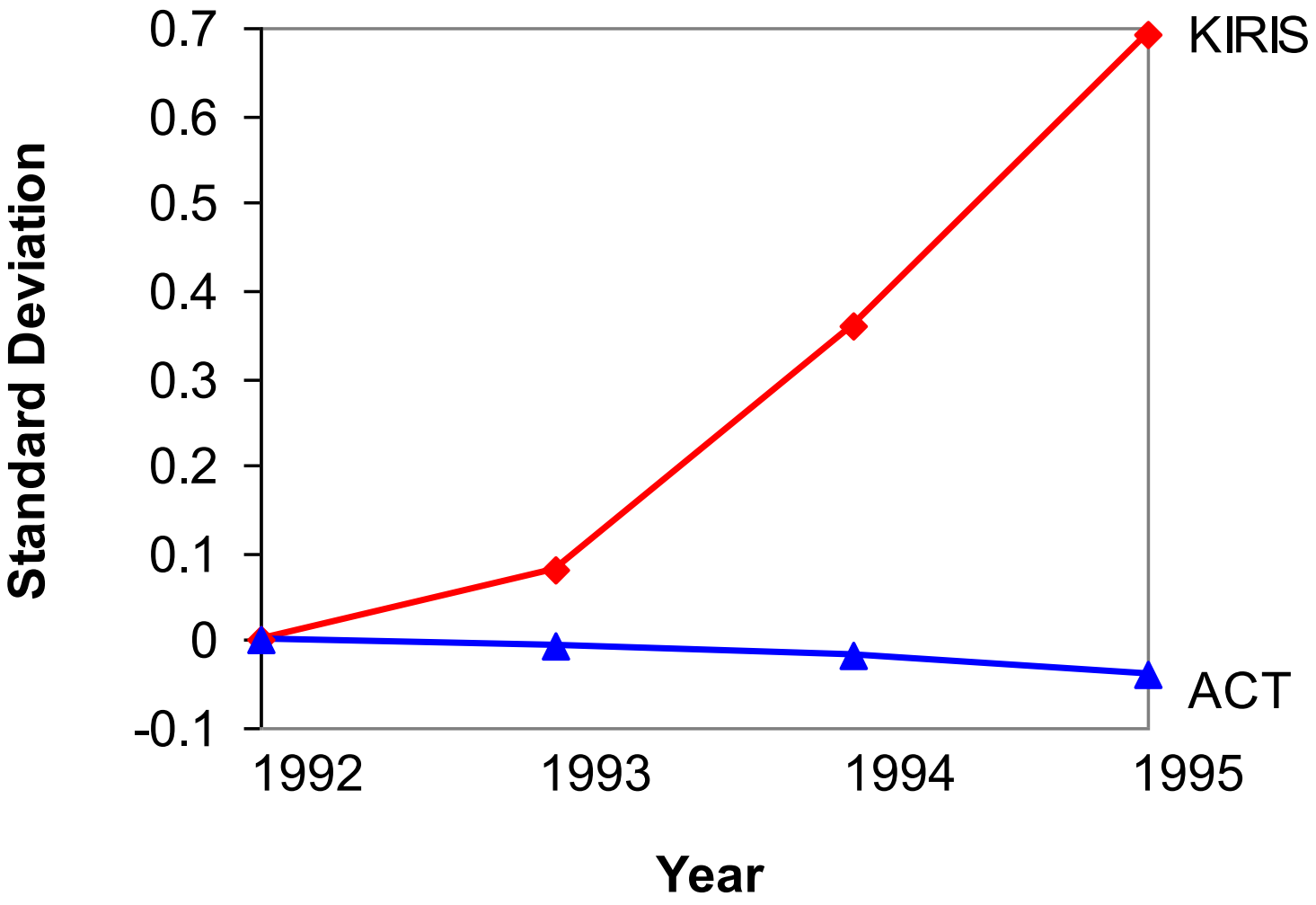
Donald T. Campbell, (1975). "Assessing the impact of planned social change." In G. M. Lyons (Ed.), *Social Research And Public Policies : The Dartmouth/OECD Conference*.

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:

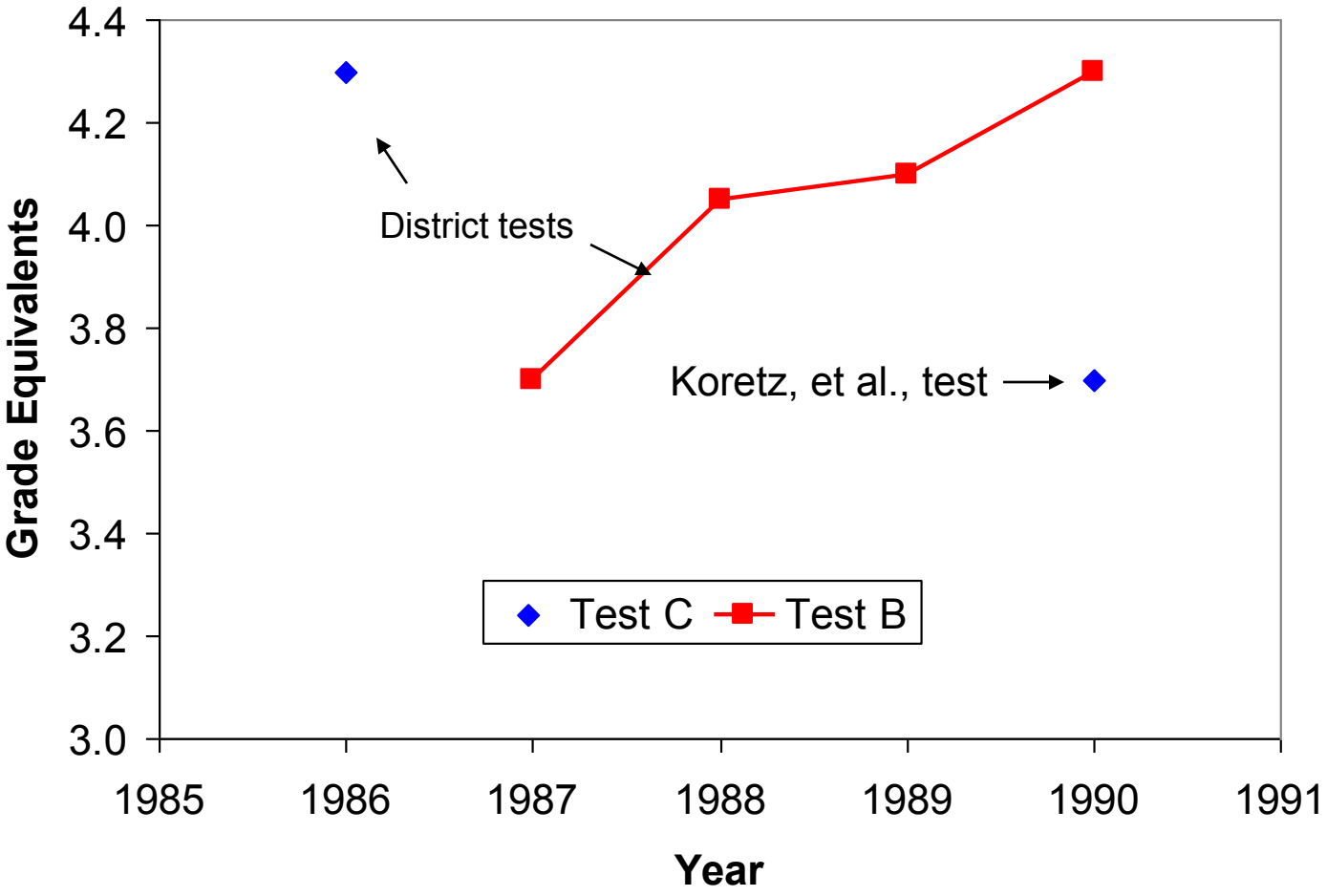
https://my.vanderbilt.edu/performanceincentives/files/2012/10/200804_Rothstein_HoldingAccount.pdf


Standard Deviation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
1992
1993
1994
1995
Year
KIRIS
ACT

Standardized mathematics gains in Kentucky, 1992-1996

	KIRIS	NAEP
Grade 4	0.61	0.17
Grade 8	0.52	0.13

Performance on coached and uncoached tests

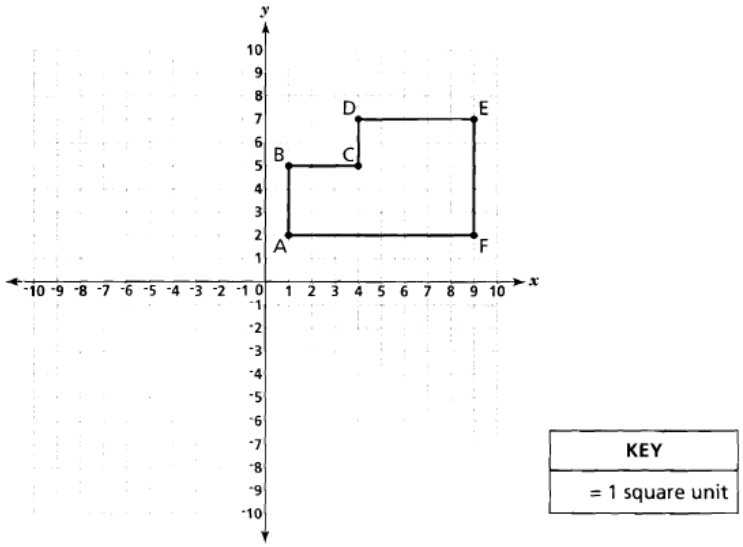


SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)

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?

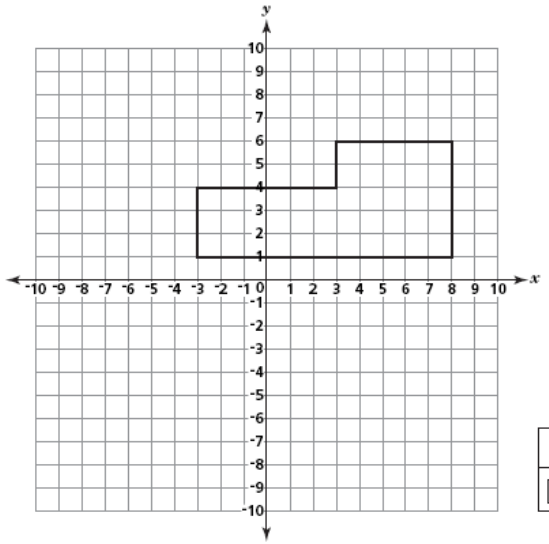
A 40

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

Samples from three word lists

A	B	C
siliculose	bath	feckless
vilipend	travel	disparage
epimysium	carpet	miniscule

New samples from three word lists

A	B	C
siliculose	bath	feckless/ parsimonious
vilipend	travel	disparage
epimysium	carpet	miniscule