Using Tests for Evaluation: Experiences from the U.S.

Daniel Koretz
Harvard Graduate School of Education

3rd ICFES International Seminar of Educational Research

1 November 2012

Problem statement

- Growing interest worldwide in using tests for evaluation and test-based accountability (TBA)
  - To monitor the performance of schools
  - To encourage improvement
  - To select and sort students
- 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
- Need to build systems that minimize these problems

What we do and do not know about high-stakes testing

- The effect on student achievement is unclear
  - Weak research designs, weaker data
  - Some evidence of inconsistent, modest effects
- Effects on educational practice are mixed
  - Some improvements
  - Some undesirable effects—bad test preparation, other “gaming”
- Scores can become severely inflated (increase much more than actual learning)

Topics

- The “sampling principle” of testing
- Evidence of 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

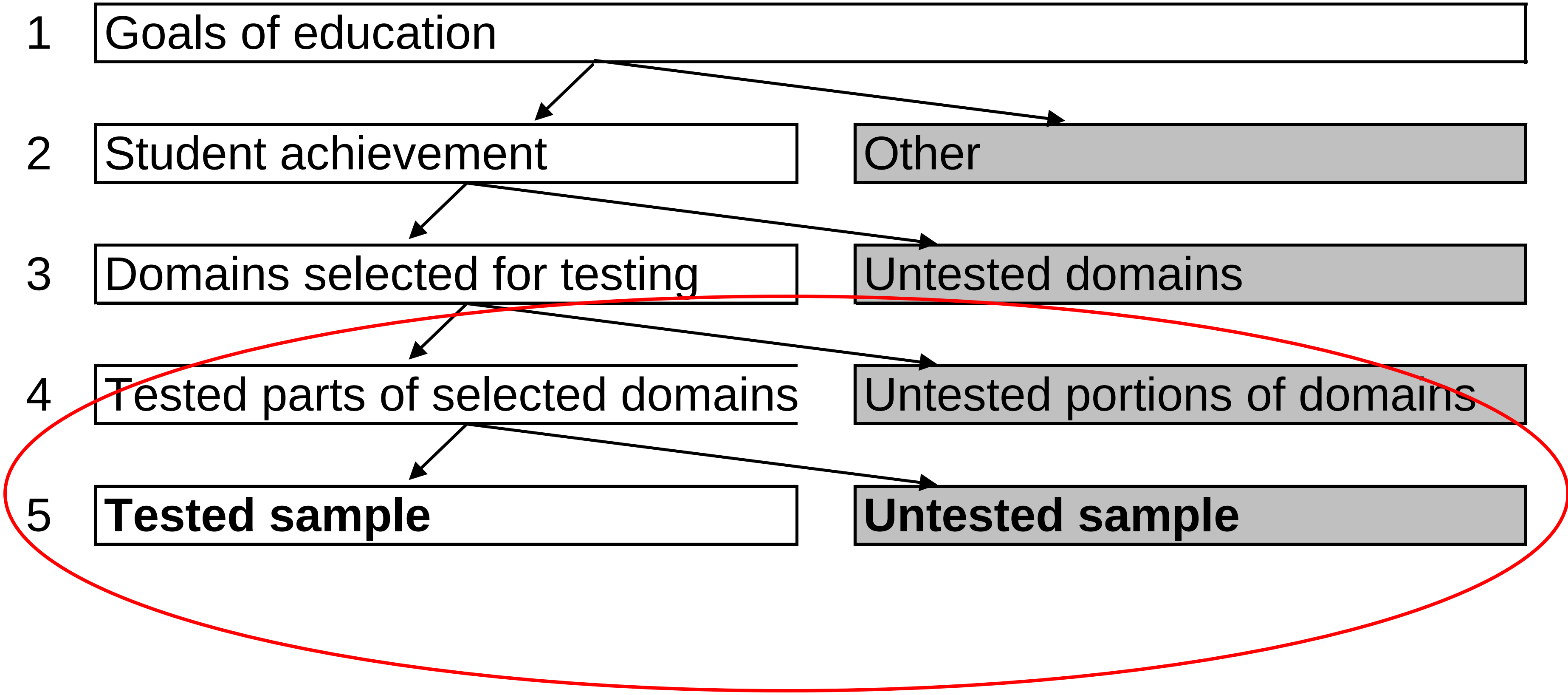
- On 3 June 2010, a poll of 2.000 people poll by Centro Nacional de Consultorí predicted 61.6% for Santos, 29.8%for Mockus

- 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?

Sampling to obtain a test



1 Goals of education

2 Student achievement | Other

3 Domains selected for testing | Untested domains

4 Tested parts of selected domains | Untested portions of domains

5 **Tested sample** | **Untested sample**

What are the consequences of incomplete sampling?

- All cases:
  - Systematically incomplete evaluation of education
- Low pressure: modest effects
  - Measurement error (uncertainty): fluctuations in scores
  - (Usually) modest differences among tests
- High pressure (accountability): very large effects
  - Incentives to focus on the tested sample, not the domain
  - Narrowed instruction, bad test preparation
  - Score inflation

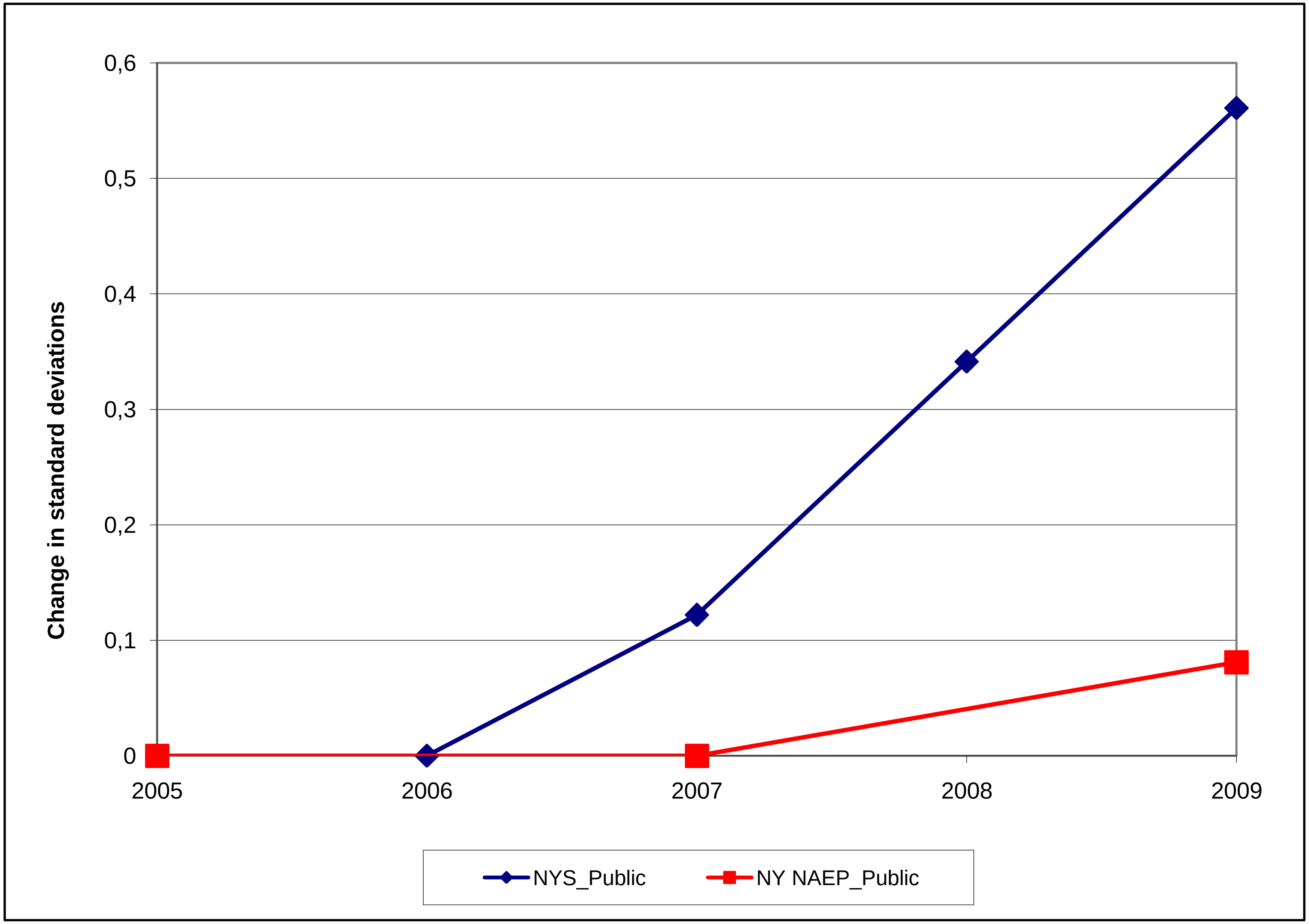
Topics

- The “sampling principle” of testing
- Evidence of score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new accountability 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 <u>must</u> generalize to other tests of the same domain
  - If a poll is accurate, other polls will show similar results

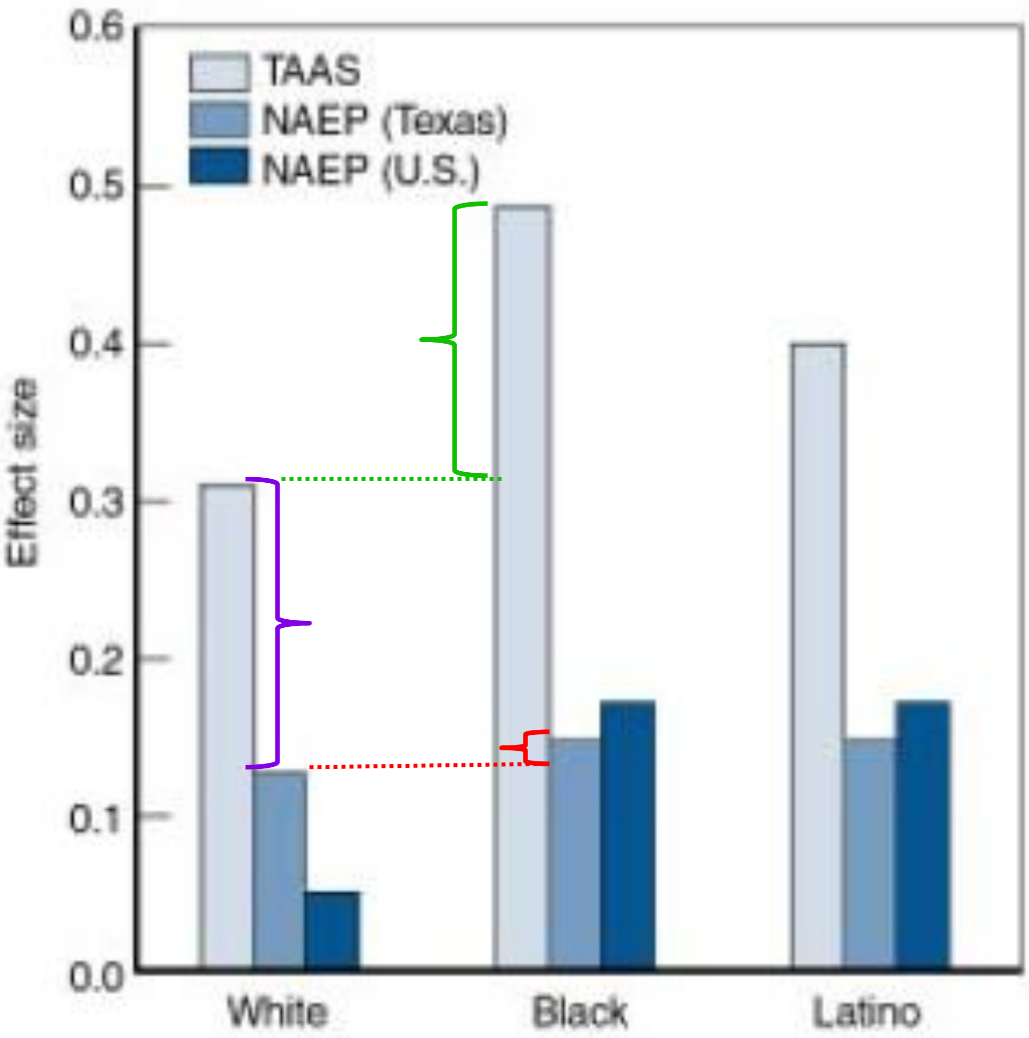
Grade 8 math score trends in New York State



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

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

One look at the "Texas miracle" (Klein, et al. 2000)



Topics

- The “sampling principle” of testing
- Evidence of score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for developing new accountability 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 test-specific gains that do not generalize beyond that test

Ways to raise scores

Teaching more

Working harder

Working more effectively

Reallocation

Coaching

Cheating

Changing who is tested

Reallocation

- Shifting instructional resources to fit the testing program
  - Within a subject
  - Between subjects
- Reallocates achievement
- Within a subject, can lead to either meaningful change or inflation
  - Inflates if material getting *decreased* emphasis is also important for the inference

Opportunities for reallocation

Recurrent and predictable patterns in the test:

- Recurrent emphasis
  - Some tested content appears more on tests than other content
- Recurrent omissions from the test

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.

Coaching

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

Opportunities for coaching

Again, recurrent patterns in the test

- Recurrent minor details of content (emphasis and omission)

- Recurrent forms of presentation
  - Item formats
  - Other aspects of presentation

- Recurrent response demands (e.g., how work is scored)

2008 item, New York grade 7 math test

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

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

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
- Evidence of score inflation
- Responses to high-stakes testing: how score inflation happens
- Implications for improving the use of tests for evaluation and accountability

What we know

- Using only a test for evaluation and accountability is not adequate
  - 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:
  - Overallimprovement is exaggerated
  - Relativeeffectiveness (for example, of schools) is estimated incorrectly

What about value-added modeling (VAM)?

- VAM is in some ways the better way to measure output
- VAM raises many additional issues, for example:
  - Large amounts of random noise
  - Unstable results from one test to another
  - Difficulty inferring true effects of teachers and schools
- VAM does <u>not</u> address the problem of score inflation

What we don’t know

- We have not identified the best types of test-based evaluation and accountability systems
  - Which programs maximize real improvements
  - Which programs minimize gaming, bad test preparation, & score inflation
- Reason: grossly inadequate research and evaluation
  - Research does show reasons for concern
  - Research has not yet evaluated better program designs

Suggestions

- Carry out ongoing monitoring and evaluation
  - Evaluate the evaluation system, not just education

- Try new test designs

- Try new designs for the larger accountability system

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

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 to experiment with new evaluation 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

Next steps: four key points

- Be cautious: take advantage of problems shown by US experience

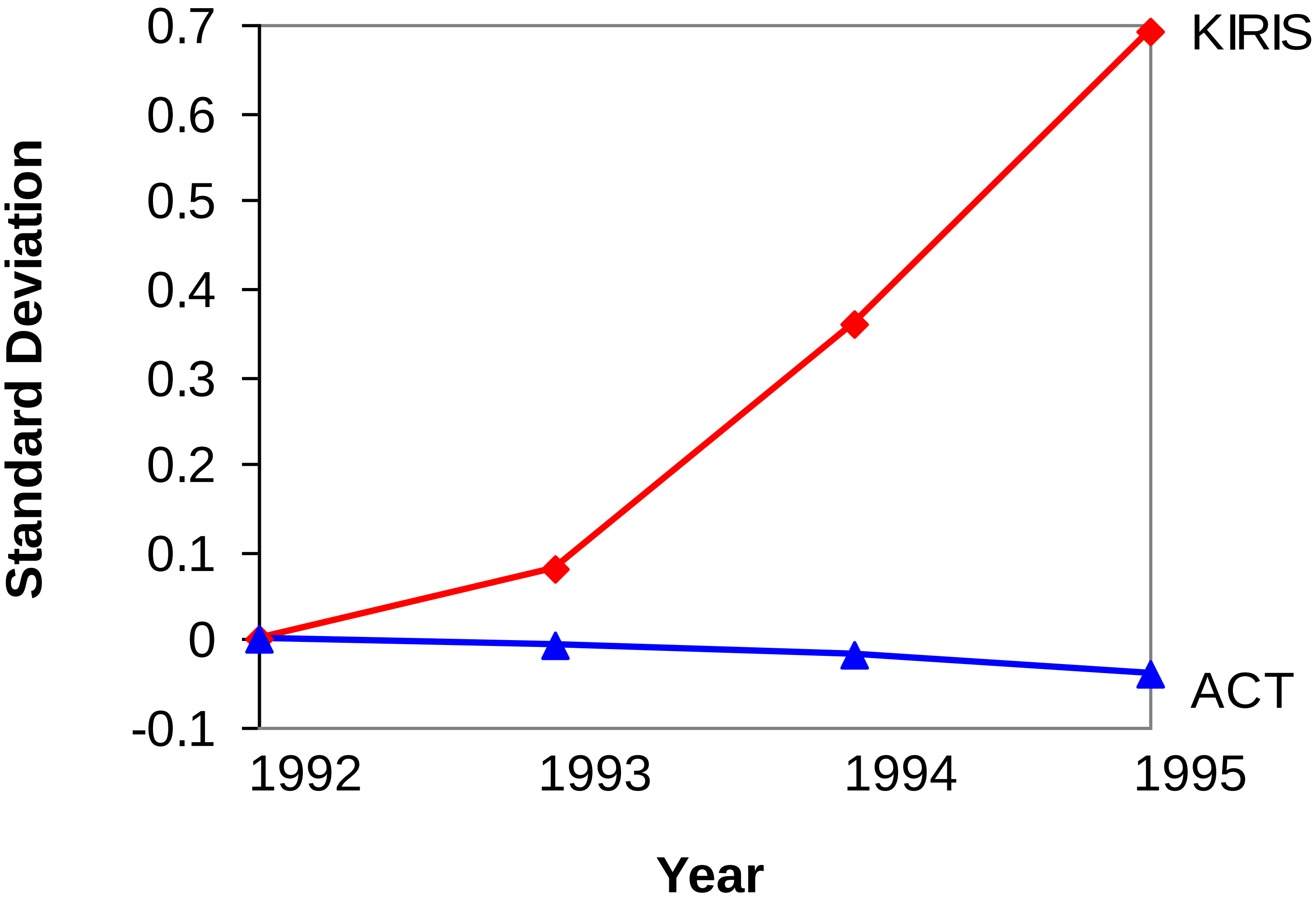
- Try a broader focus: not just test scores

- Monitor and evaluate the system routinely, and be prepared to modify testing and evaluation programs

- Participate in forward-looking research and development

Supplementary slides

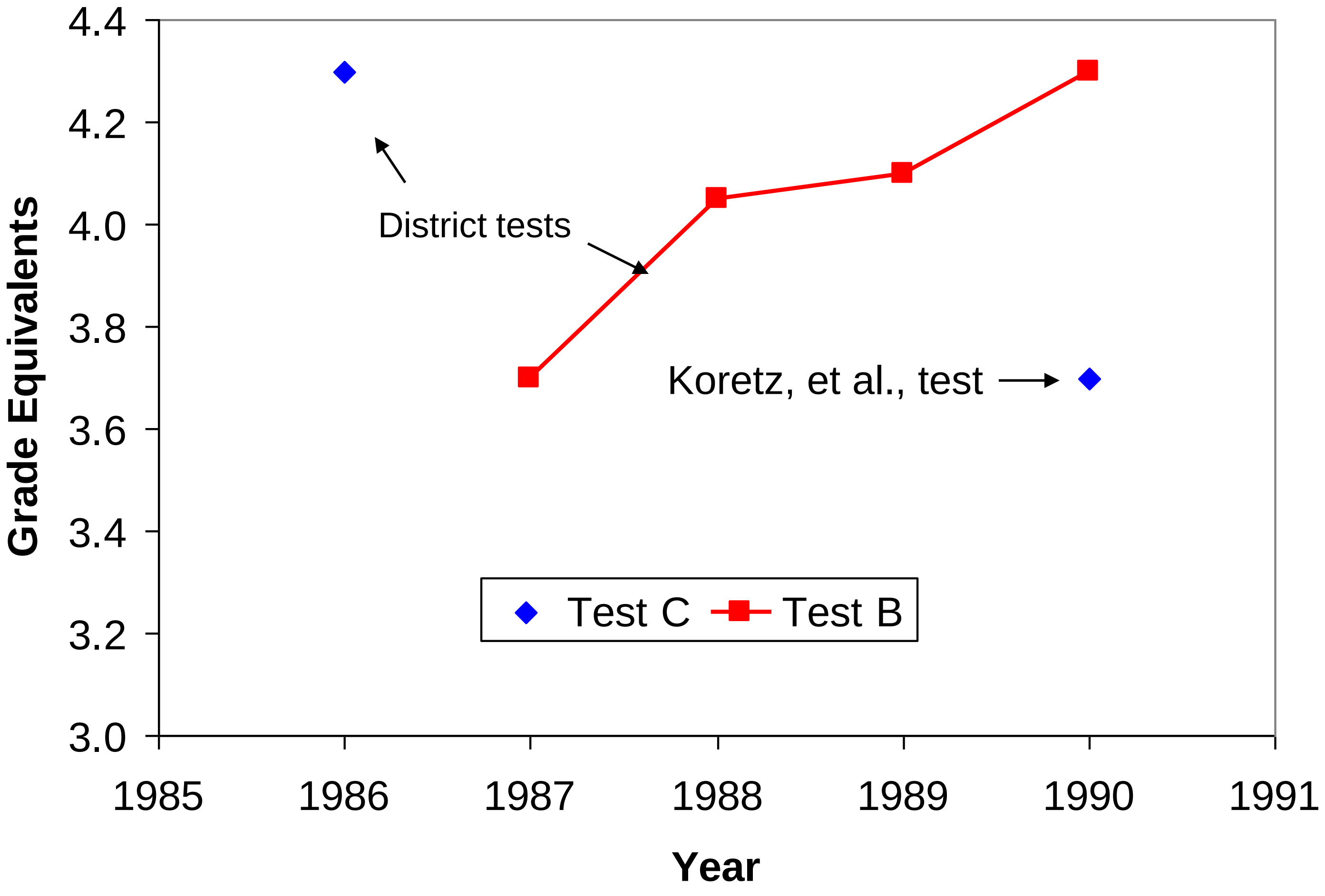
Math trends, KIRIS and ACT



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)

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

“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:

http://www.performanceincentives.org/data/files/directory/ConferencePapersNews/Rothstein.pdf