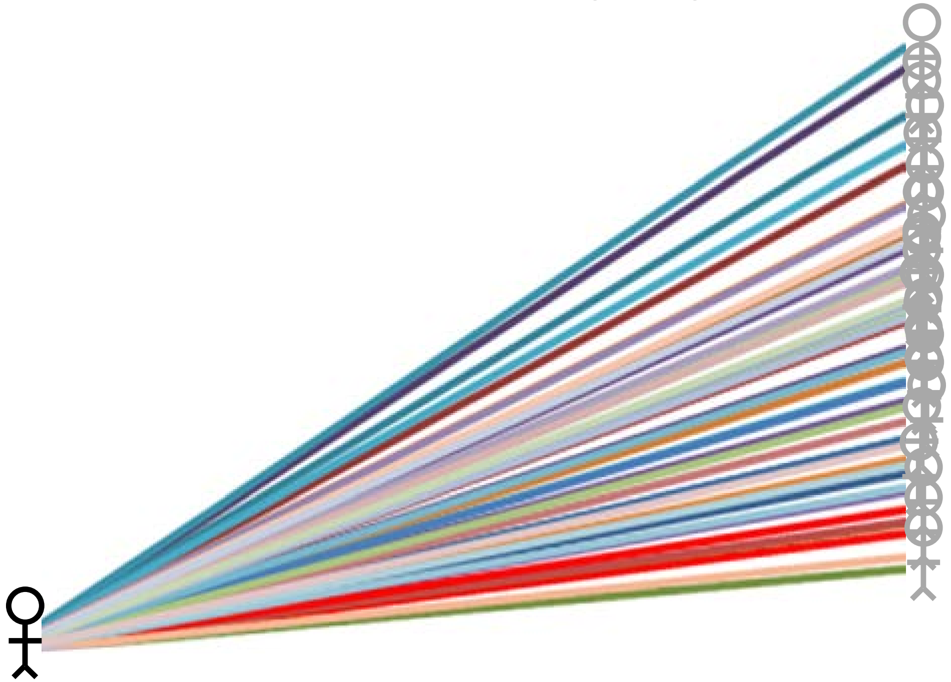
HARVARD
GRADUATE SCHOOL OF EDUCATION

Design principles for assessment-based accountability systems



Andrew Ho, Harvard Graduate School of Education
Invited Testimony
Texas Commission on Next Generation Assessments and Accountability
Austin, Texas, January 20, 2016

10 principles for test-based accountability systems

1. Encourage inclusion.
2. Refresh assessments yearly.
3. Use multiple measures.
4. Emphasize school improvement; downplay school rankings.
5. Emphasize student growth; also emphasize student proficiency.
6. Factor score precision into high-stakes decisions.
7. Budget for responses to unintended consequences.
8. Answer the question, “So what can I do about it?”
9. Anchor scales: What does a “B” or a “50” mean?
10. Increase research capacity.

Principles 1-7 adapted from Linn (2001)

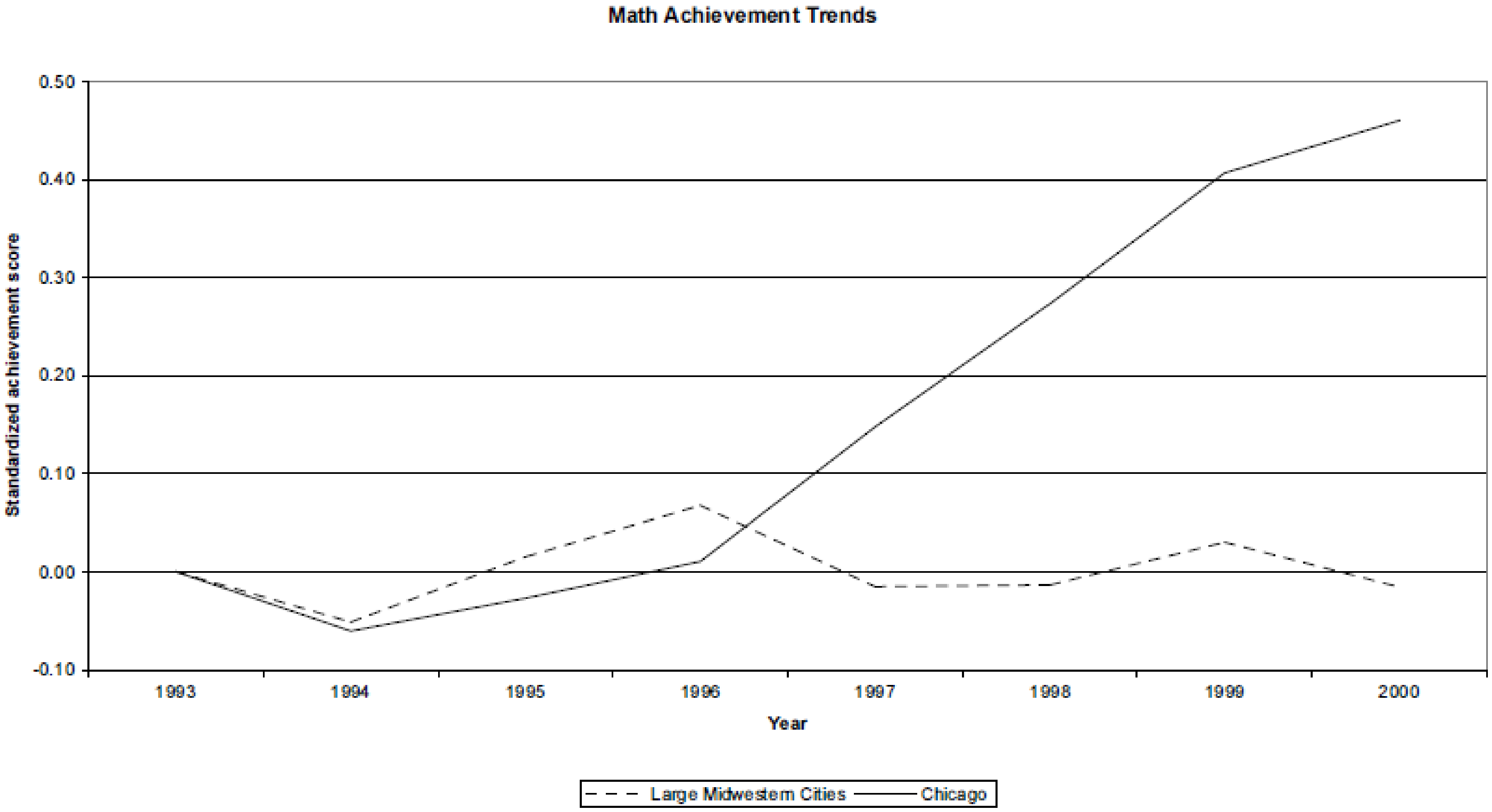
1) Encourage inclusion (Linn, 2001)

Provide safeguards against selective exclusion of students from assessments.

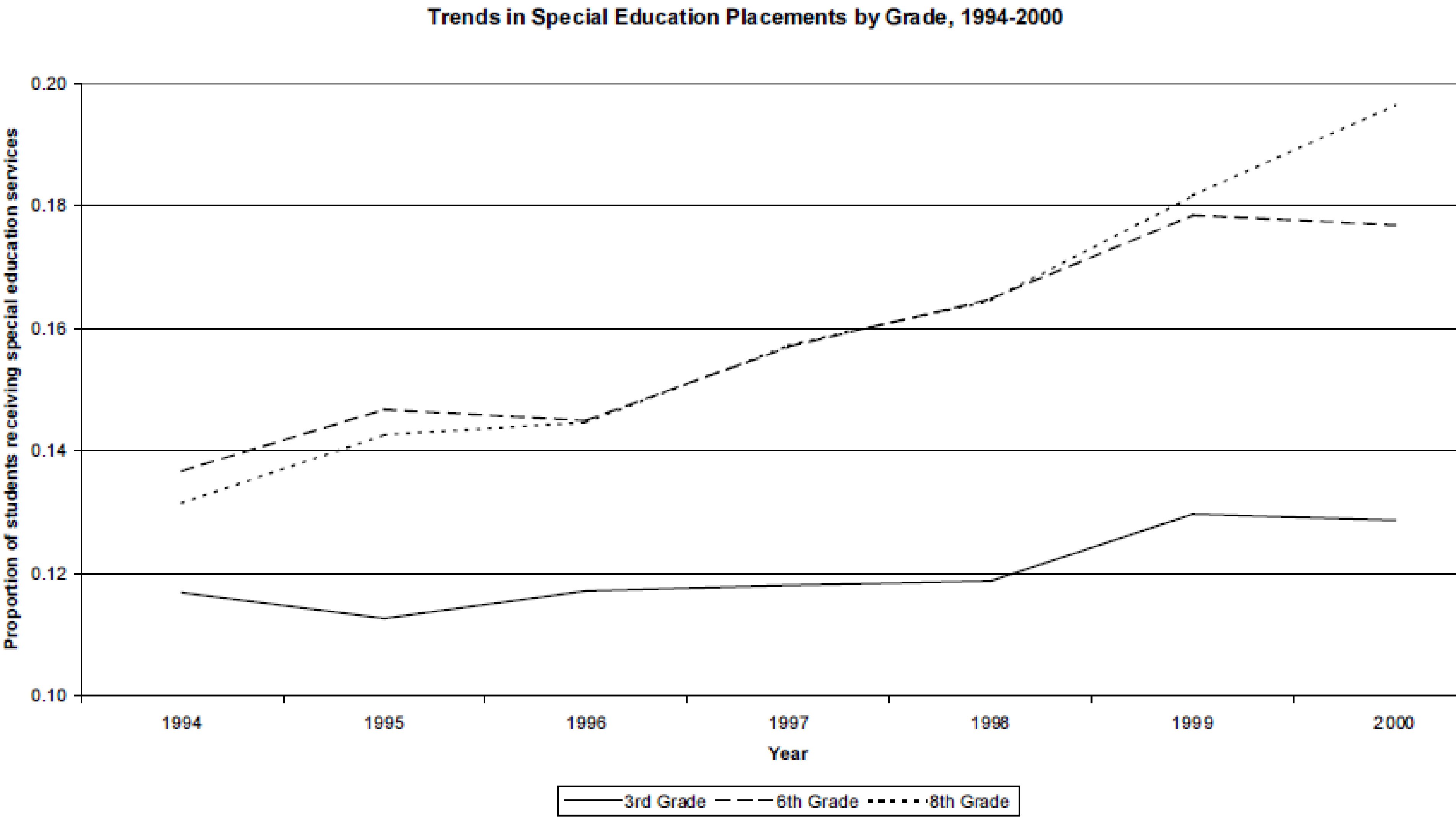
Linn (2001)

1) Encourage inclusion (Jacob, 2005)

Figure 3: Achievement Trends in Chicago versus Other Large, Urban School Districts the Midwest, 1990-2000



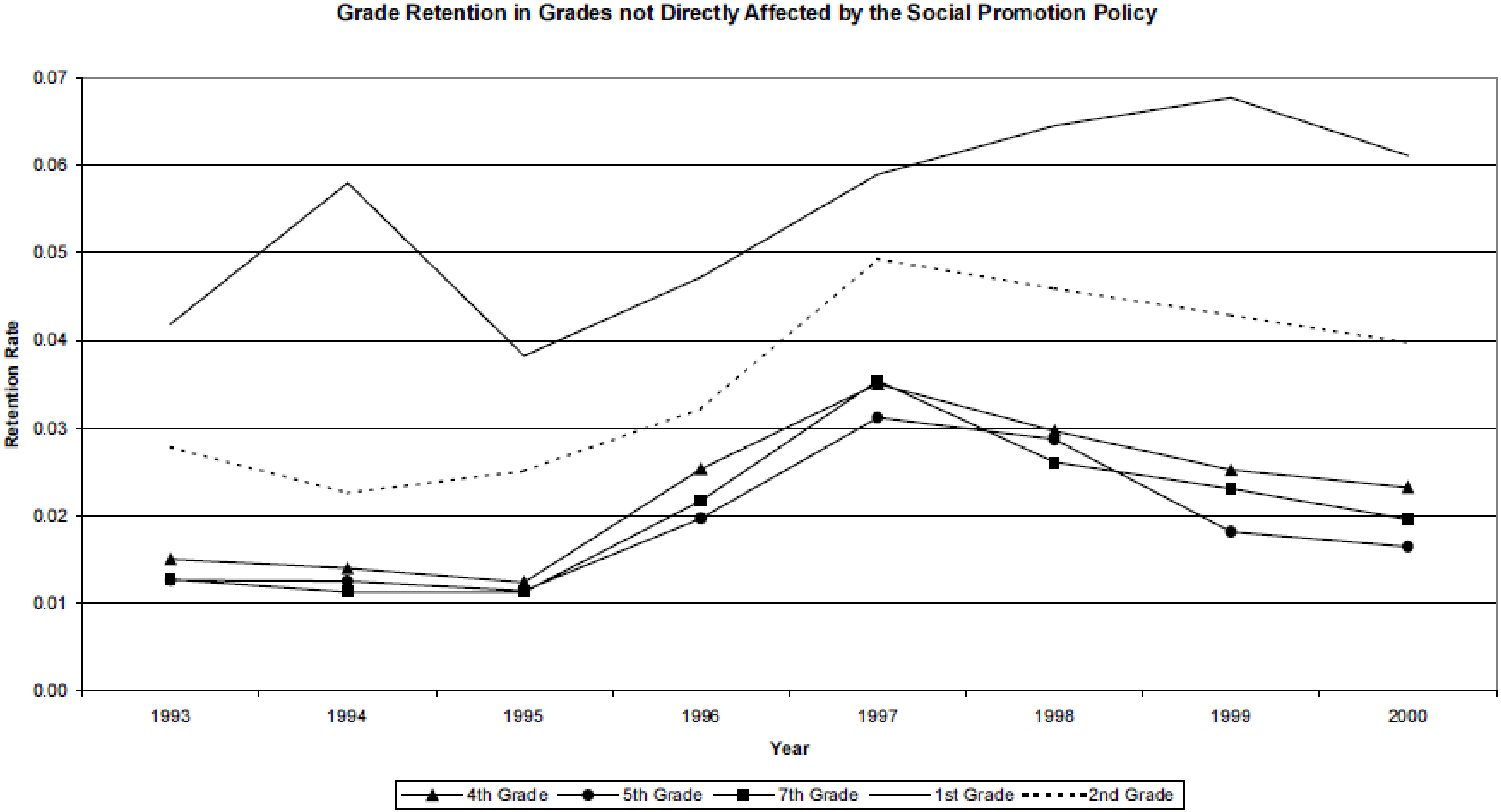
1) Encourage inclusion (Jacob, 2005)



Notes: The sample includes only first-time, non-bilingual students.

1) Encourage inclusion (Jacob, 2005)

Figure 6: Trends in Grade Retention



1) Encourage inclusion

Policy tools (each with pros and cons) include:

- Participation requirements (ESSA: 95%)
- Limiting alternative assessment participation (ESSA: 1%)
- Subgroup reporting
  - Lower minimum subgroup size (TX: 25*)
  - No super subgroups (ESSA)
- Track all participation and classification rates over time.
- Budget for unanticipated unintended responses.
- Ensure that assessment provides useful, relevant information and diagnoses achievement disparities.

2) Refresh assessment items yearly (Linn, 2001)

Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years.

Linn (2001)

2) Refresh assessment items yearly (Koretz & Barron, 1998)

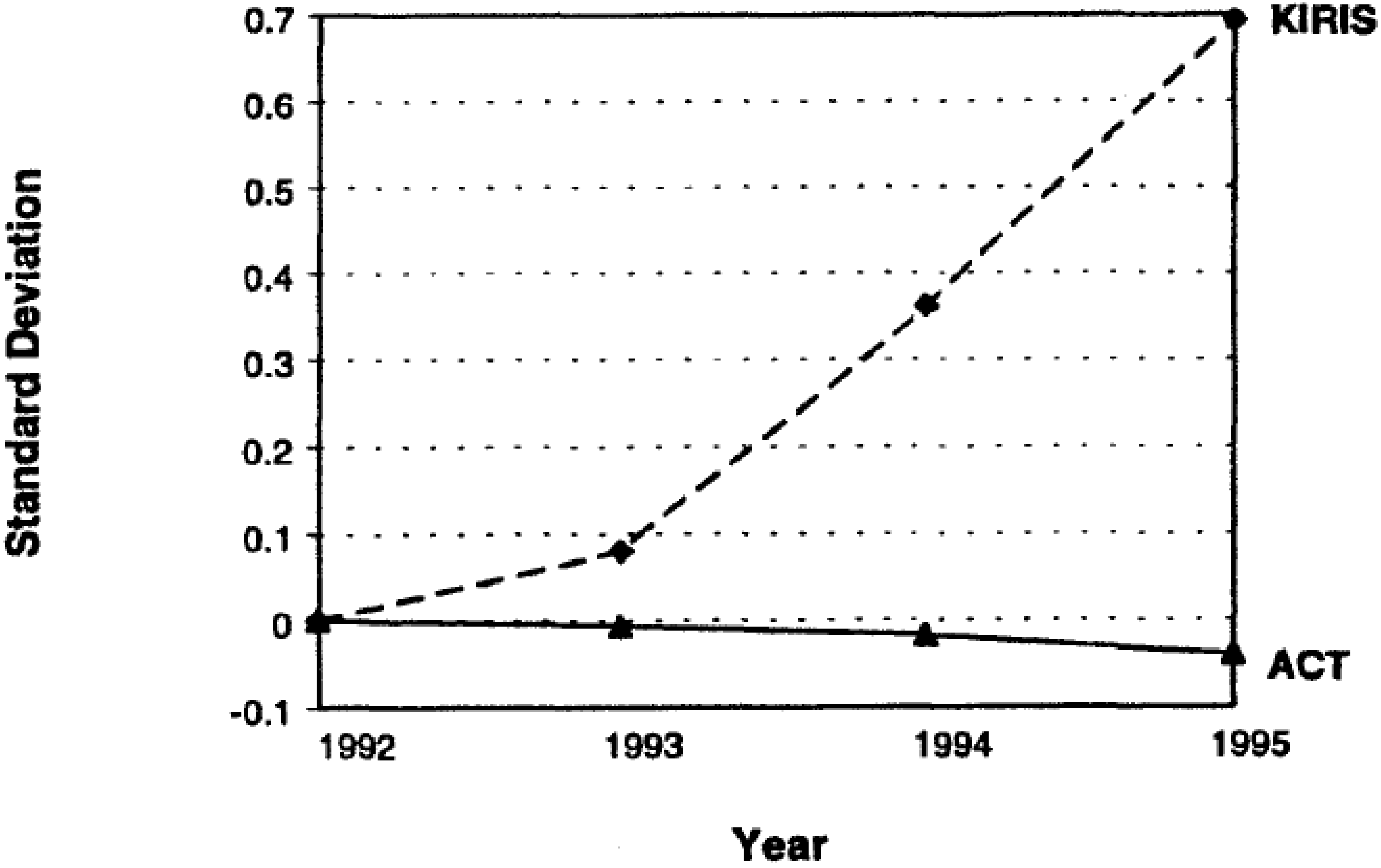


Figure 25—Standardized Changes in ACT and KIRIS Mathematics

2) Refresh assessment items yearly (Koretz & Barron, 1998)

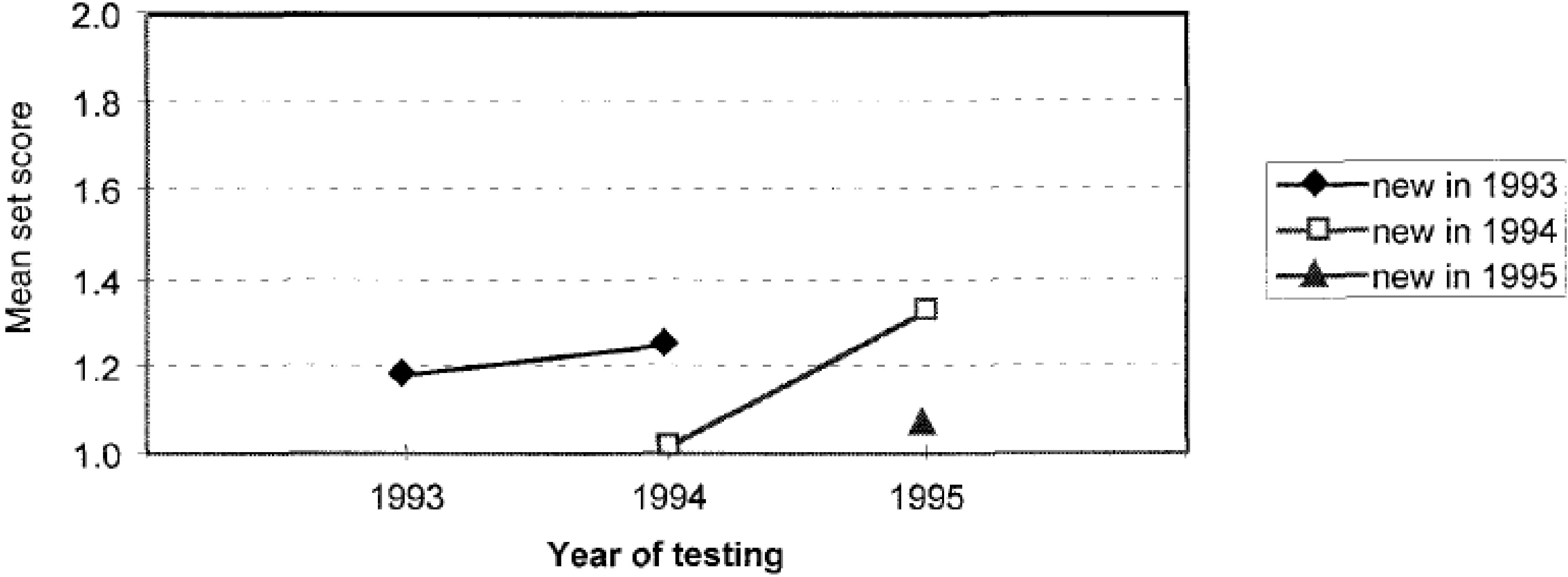


Figure 30—Sawtooth Pattern: KIRIS Grade 8 Mathematics, All Items

2) Refresh assessment items yearly

Policy tools (each with pros and cons):

- Invest significantly in assessment and item development
- Emphasize trends over time as a contribution of the system.
- Budget for the significant costs of maintaining comparable assessments over time

3) Use multiple measures (Linn, 2001)

Don’t put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.

Linn (2001)

3) Use multiple measures (Ho, 2007)

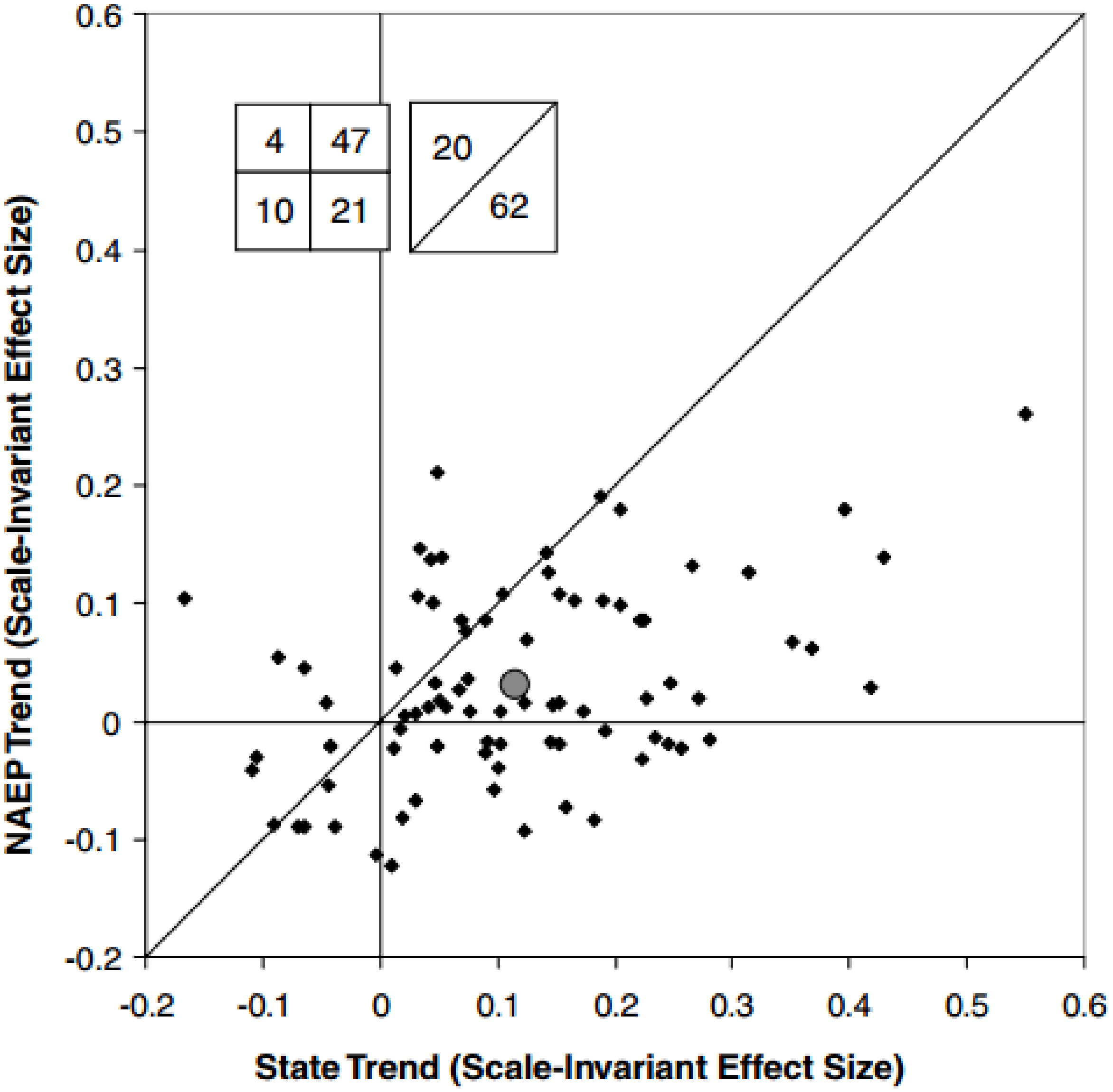


FIGURE 4. NAEP versus State Score Trend Discrepancies; All 82 State-Subject-Grade Combinations.

3) Use multiple measures (B&M Gates Found., 2013)

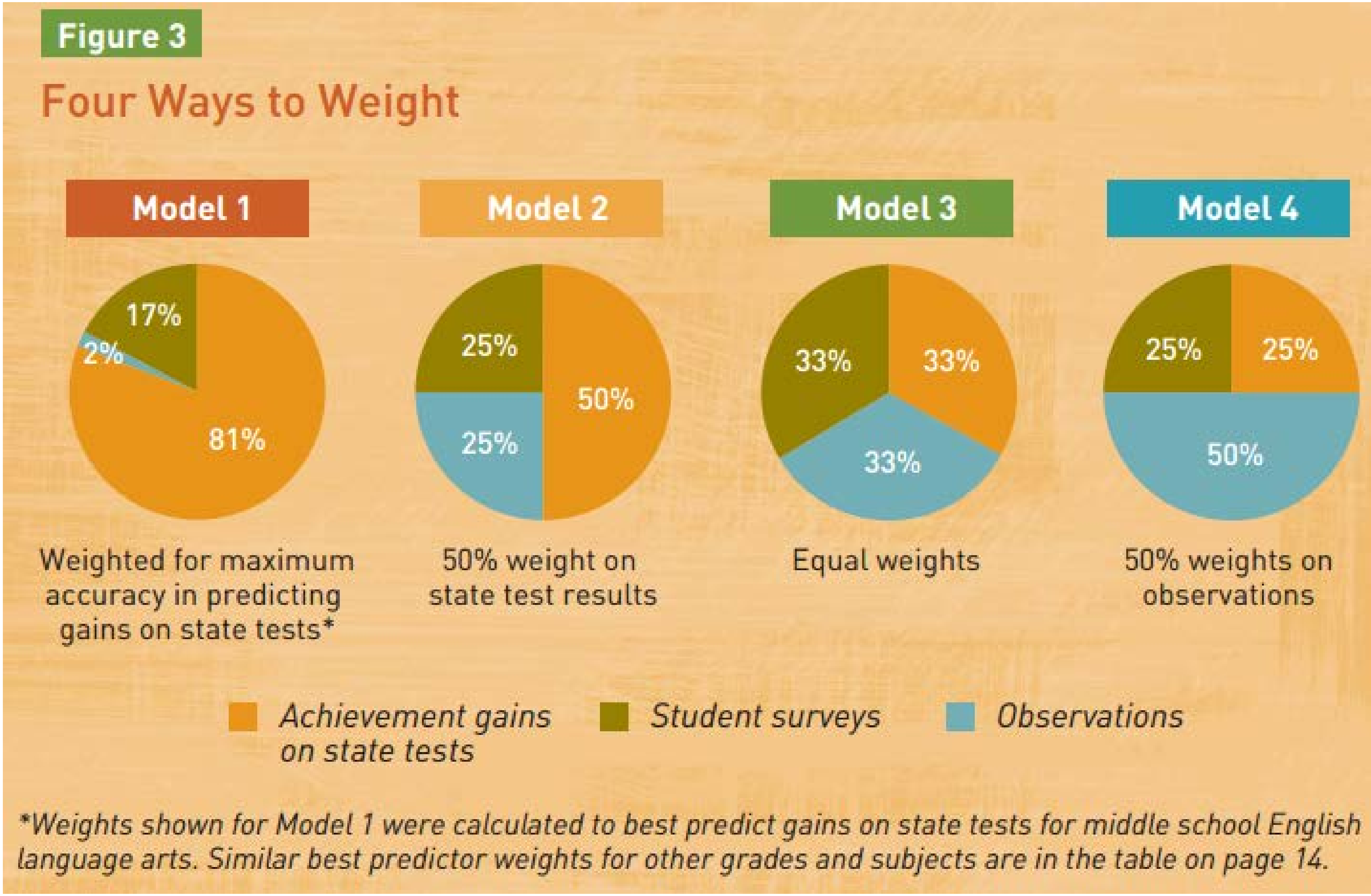


Figure 3

Four Ways to Weight

Model 1: Weighted for maximum accuracy in predicting gains on state tests*

Model 2: 50% weight on state test results

Model 3: Equal weights

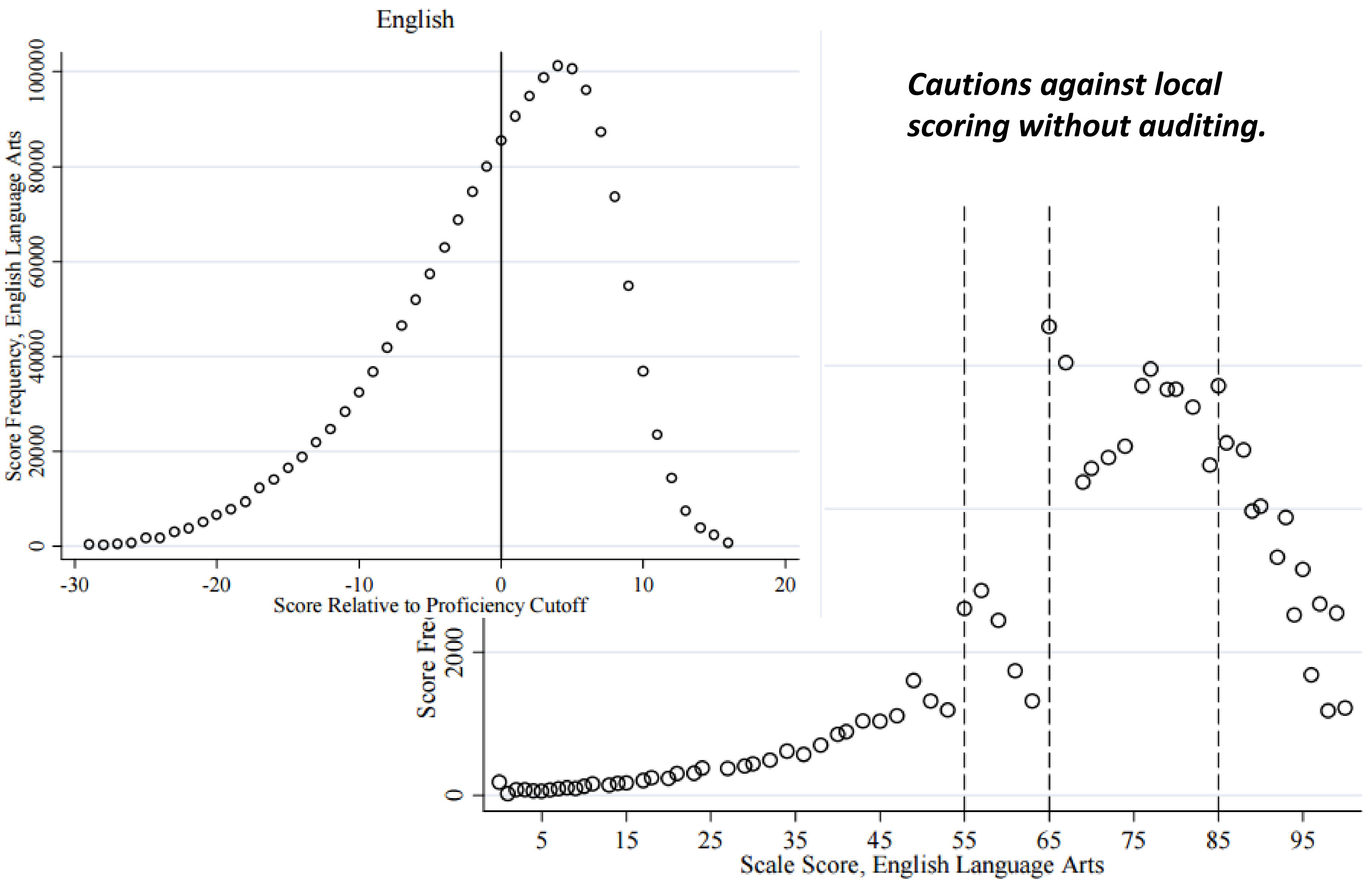
Model 4: 50% weights on observations

🟧 *Achievement gains on state tests* 🟫 *Student surveys* 🟦 *Observations*

Weights shown for Model 1 were calculated to best predict gains on state tests for middle school English language arts. Similar best predictor weights for other grades and subjects are in the table on page 14.

3) Use multiple measures (Dee, Jacob, McCrary, Rockoff, 2011)

Figure 3: Frequency Distributions for Centrally Graded Exams in Grades 3 to 8



Cautions against local scoring without auditing.

3) Use multiple measures

Policy tools

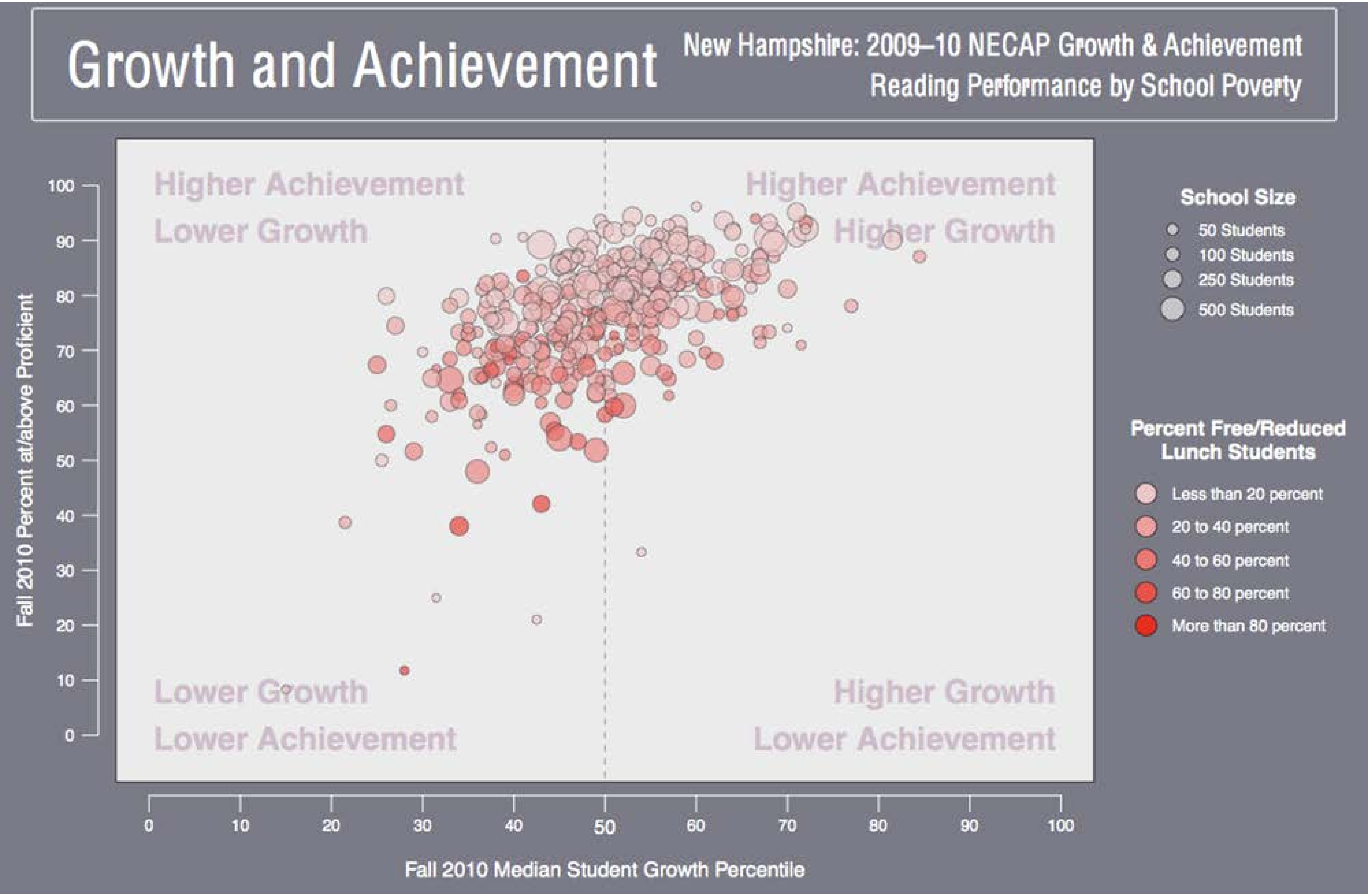
- Dashboards
- Local assessments (Student Learning Objectives)
- Index systems (e.g., TX)
- Assign higher weights to more precise measures
- Lower stakes

4) School improvement over school rankings (Linn, 2001)

Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.

Linn (2001)

4) School improvement over school rankings (NECAP, 2010)



4) School improvement over school rankings

Example (the parable of the 10-9 and 1-7 schools):

- Which school would you rather send your child to, a school that goes from a 10 to a 9, or a school that goes from a 1 to a 7? Which school would you rather laud, or sanction? [What is a 10?]

Policy tools

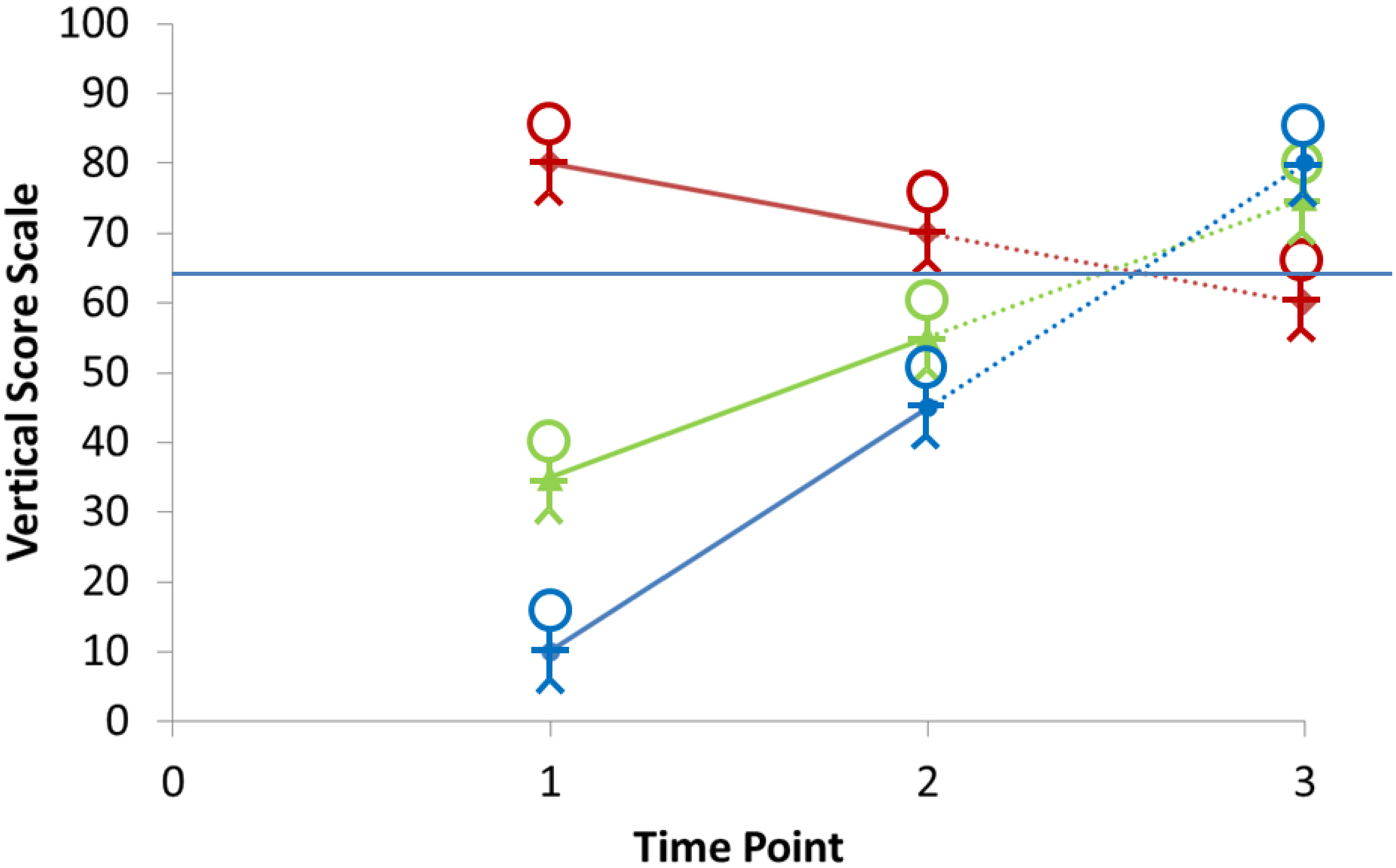
- Growth metrics (e.g., Texas)
- Score report design

5) Emphasize student growth… and proficiency (Linn, 2001)

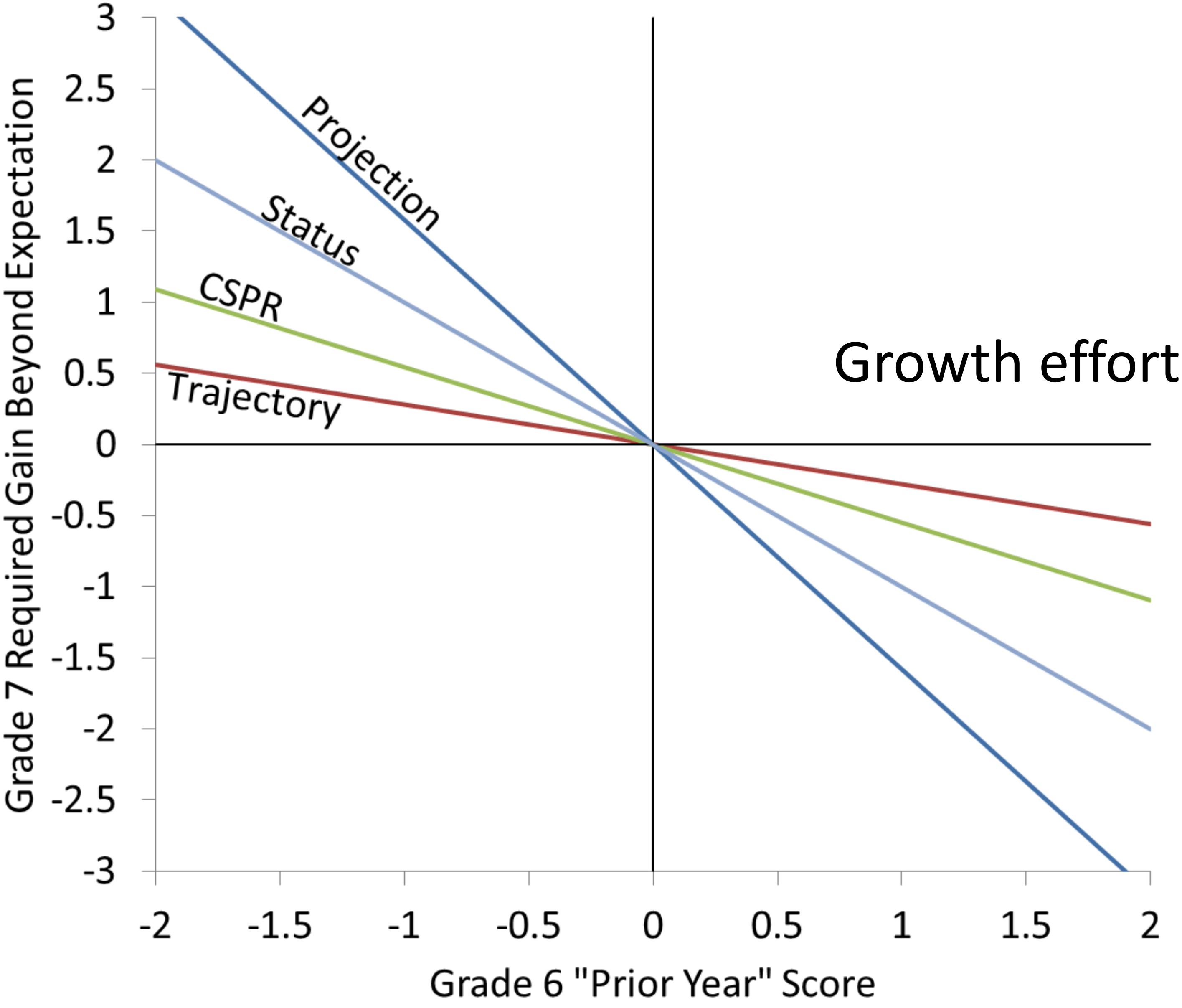
Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement while status guards against institutionalizing low expectations for those same students and schools.

Linn (2001)

5) Emphasize student growth… and proficiency (Ho, 2014)

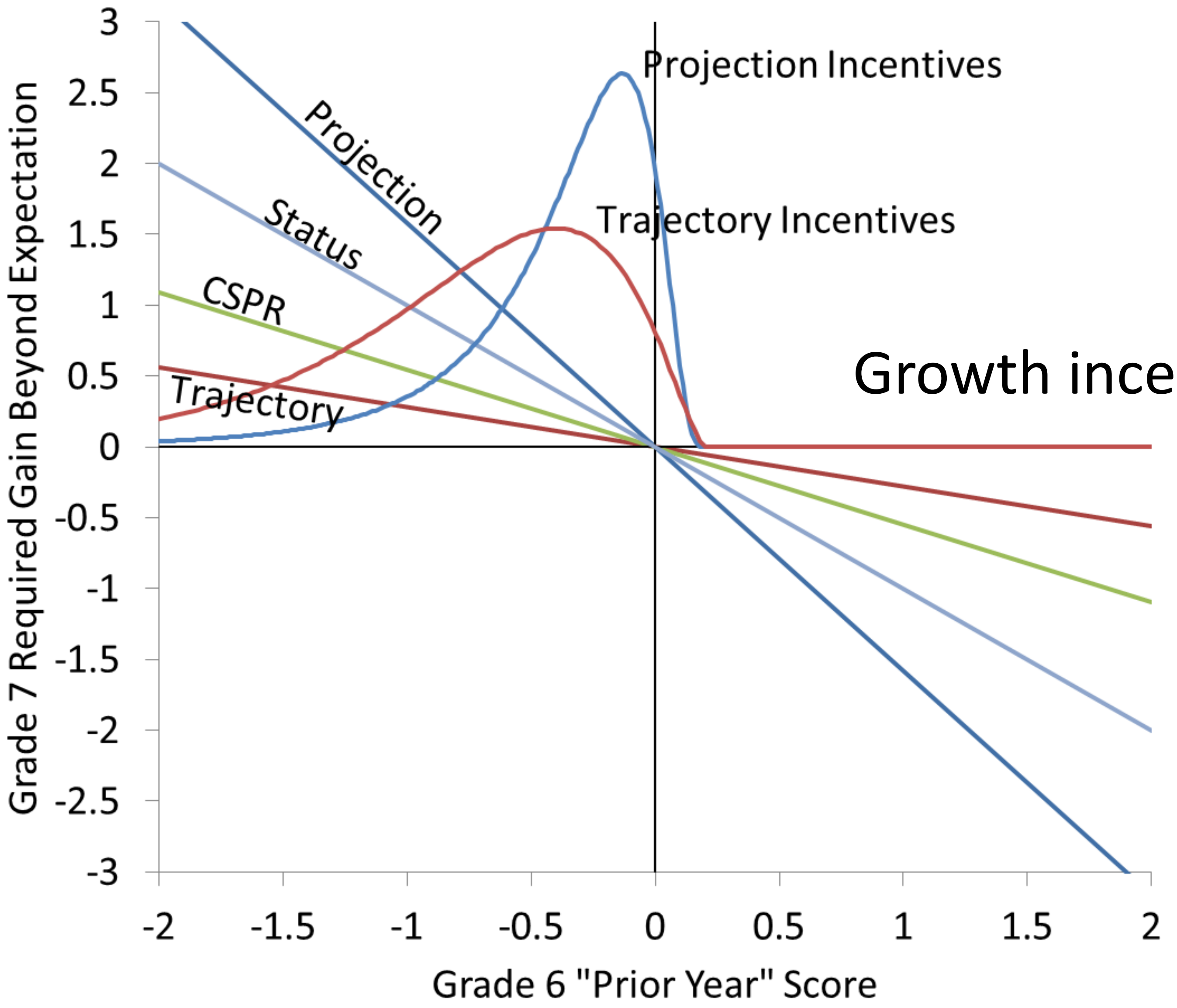


5) Emphasize student growth… and proficiency (Ho, 2014)



Growth effort map

5) Emphasize student growth… and proficiency (Ho, 2014)



Growth incentive map

5) Emphasize student growth… and proficiency

Example (revisiting the parable of the 10-9 and 1-7 schools):

- Which school would you rather send your child to, a school that takes 10s and transforms them to 9s, or a school that takes 1s and transforms them to 7s? Which school would you rather laud, or sanction?

Policy tools

- Growth metrics (e.g., Texas)
- Status metrics (college readiness benchmarks)
- Lower stakes
- Growth incentive maps

10 principles for test-based accountability systems

1. Encourage inclusion.
2. Refresh assessments yearly.
3. Use multiple measures.
4. Emphasize school improvement; downplay school rankings.
5. Emphasize student growth; also emphasize student proficiency.
6. Factor score precision into high-stakes decisions.
7. Budget for responses to unintended consequences.
8. Answer the question, "So what can I do about it?"
9. Anchor scales: What does a "B" or a "50" mean?
10. Increase research capacity.

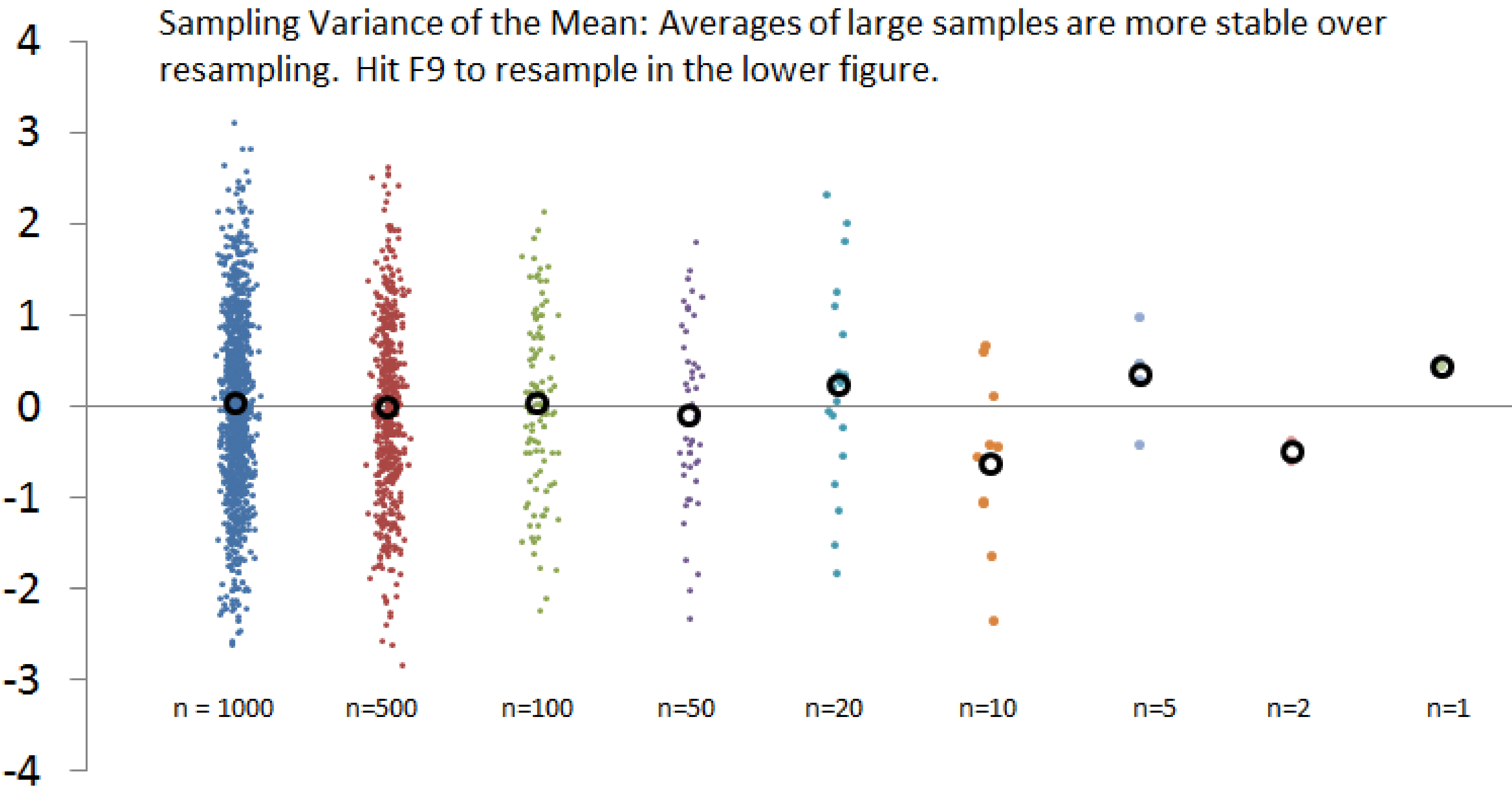
Principles 1-7 adapted from Linn (2001)

6) Factor score precision into high-stakes decisions (Linn, 2001)

Recognize, evaluate, and report the degree of uncertainty in the reported results.

Linn (2001)

6) Factor score precision into high-stakes decisions

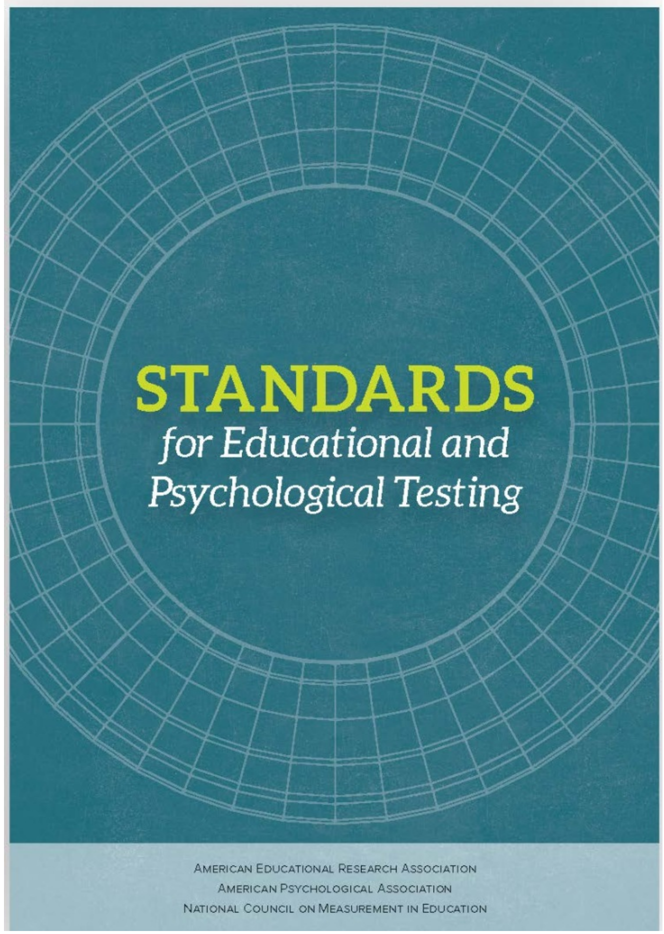


Sampling Variance of the Mean: Averages of large samples are more stable over resampling. Hit F9 to resample in the lower figure.

6) Factor score precision into high-stakes decisions

(AERA/APA/NCME Standards, 2014):

- 12.18: “score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error…”
- But also, 12.15: “Individuals who interpret the test results [should] be qualified to do so or be assisted by and consult with persons who are so qualified.”



STANDARDS

for Educational and Psychological Testing

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

6) Factor score precision into high-stakes decisions

Policy tools

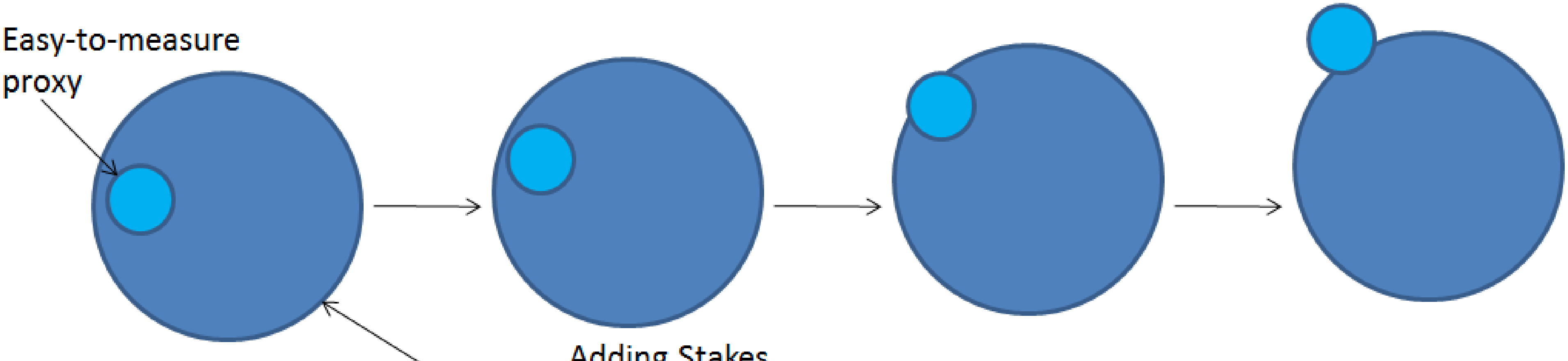
- Add standard errors and error bars to reports
- Average over measures and over time
- Adjust by confidence intervals
- Report precision-adjusted scores.

7) Budget for responses to unintended consequences (Linn, 2001)

Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

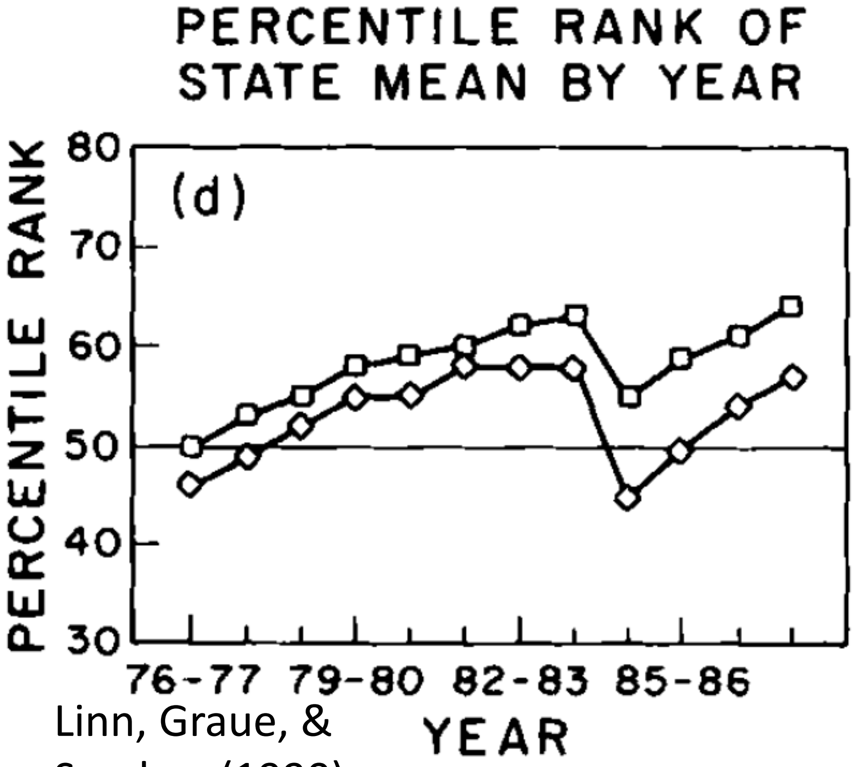
Linn (2001)

7) Budget for responses to unintended consequences



Example (AERA/APA/NCME Standards, 2014):

- 12.1: "It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible."



Linn, Graue, & Sanders (1990)

7) Budget for responses to unintended consequences

Policy tools

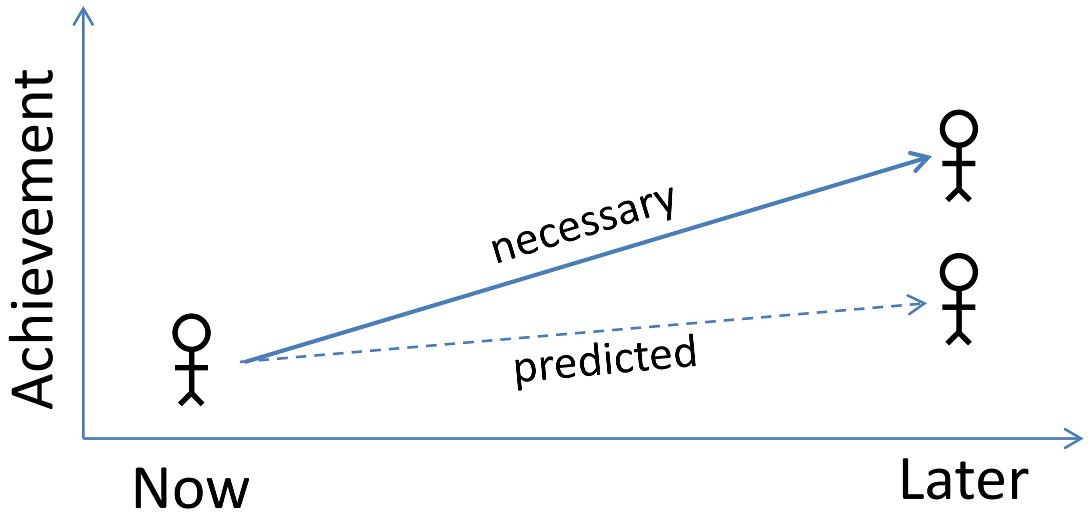
- Invest in data collection and research infrastructure.
- Research partnerships with independent evaluators.
- Encourage nimble, dynamic frameworks.
- Ongoing surveys to assess trends.
- Timed feedback loops to revisit policy features based on evidence collected by that time.

8) “So what can I do about it?” (Ho, 2014)

Accountability systems should answer two questions well:

1) Should I be worried?
2) If so, what can I do about it?

8) “So what can I do about it?” (Ho, 2014)



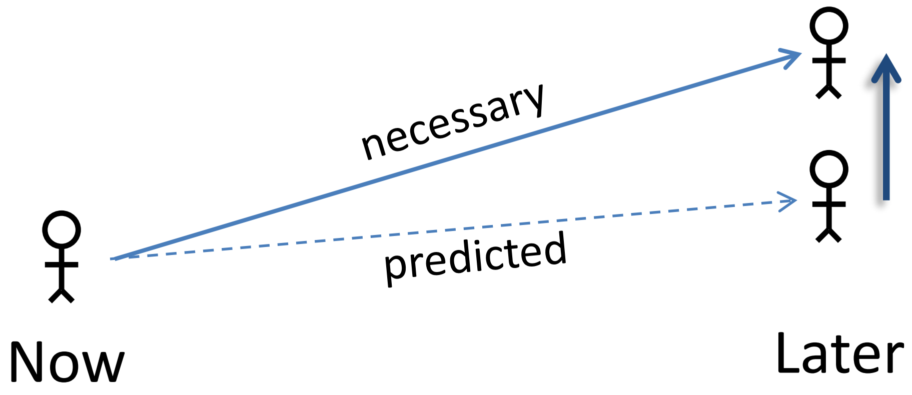
Student growth predictions should be:

a) Accurate.

b) Ultimately, incorrect.

c) Both a) and b).

8) “So what can I do about it?” (Ho, 2014)



What theories, practices, tools, policies, incentives, and interventions will lengthen ***this*** arrow?

Remember: The prediction is valid if it is ultimately wrong.

8) “So what can I do about it?”

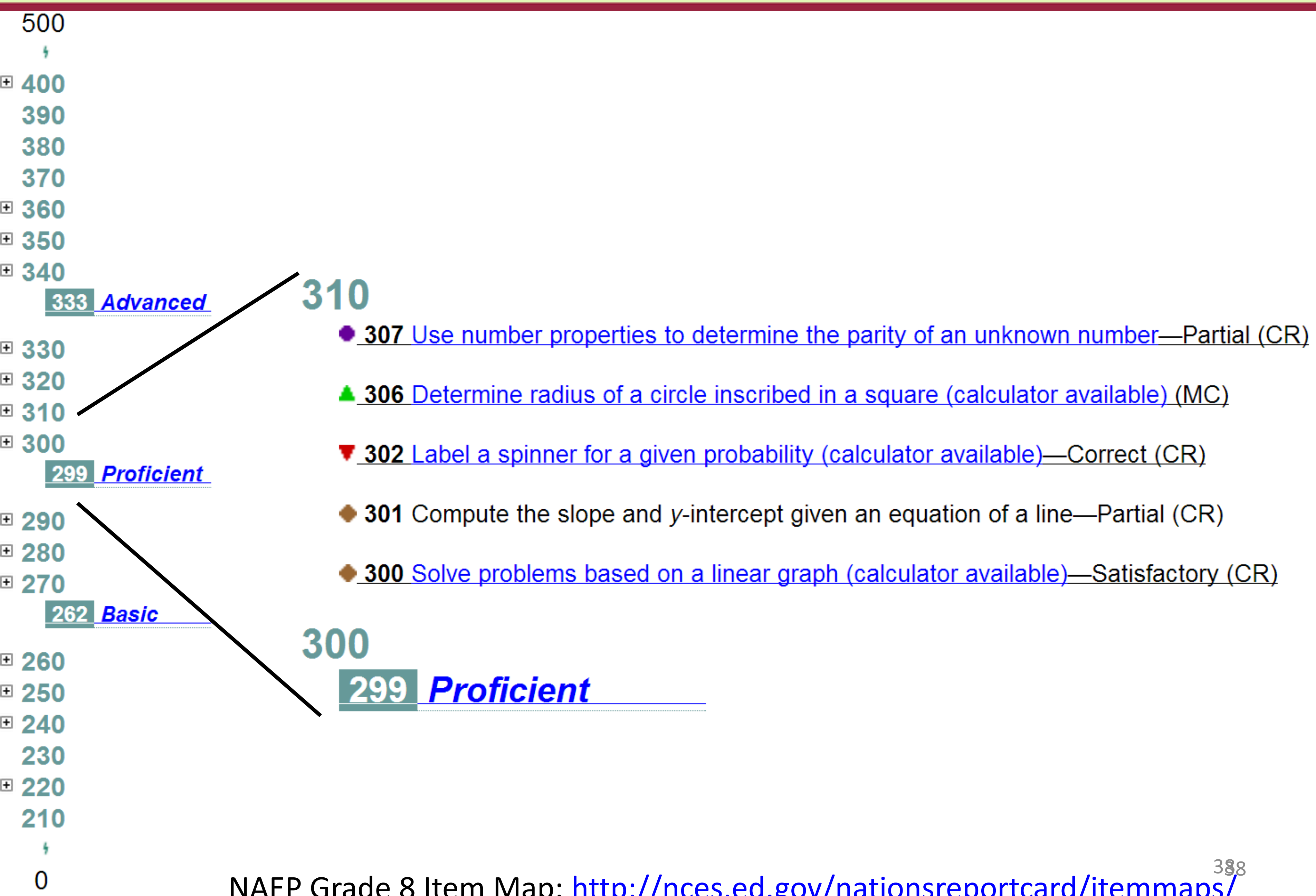
Policy tools

- Clear, timely, relevant score reporting.
- Survey stakeholders for questions they actually ask, that they would like answers to.
- Emphasize formative and diagnostic feedback
- Lower stakes

9) Anchor scales with norms and criteria (Ho)

Anchor scale points (A-F, 0-100) with explicit descriptions, including both normative (relative) and criterion (absolute) information.

9) Anchor scales with norms and criteria



500

400

390

380

370

360

350

340

333 *Advanced*

330

320

310

300

299 *Proficient*

290

280

270

262 *Basic*

260

250

240

230

220

210

0

310

● **307** Use number properties to determine the parity of an unknown number—Partial (CR)

▲ **306** Determine radius of a circle inscribed in a square (calculator available) (MC)

▼ **302** Label a spinner for a given probability (calculator available)—Correct (CR)

◆ **301** Compute the slope and *y*-intercept given an equation of a line—Partial (CR)

◆ **300** Solve problems based on a linear graph (calculator available)—Satisfactory (CR)

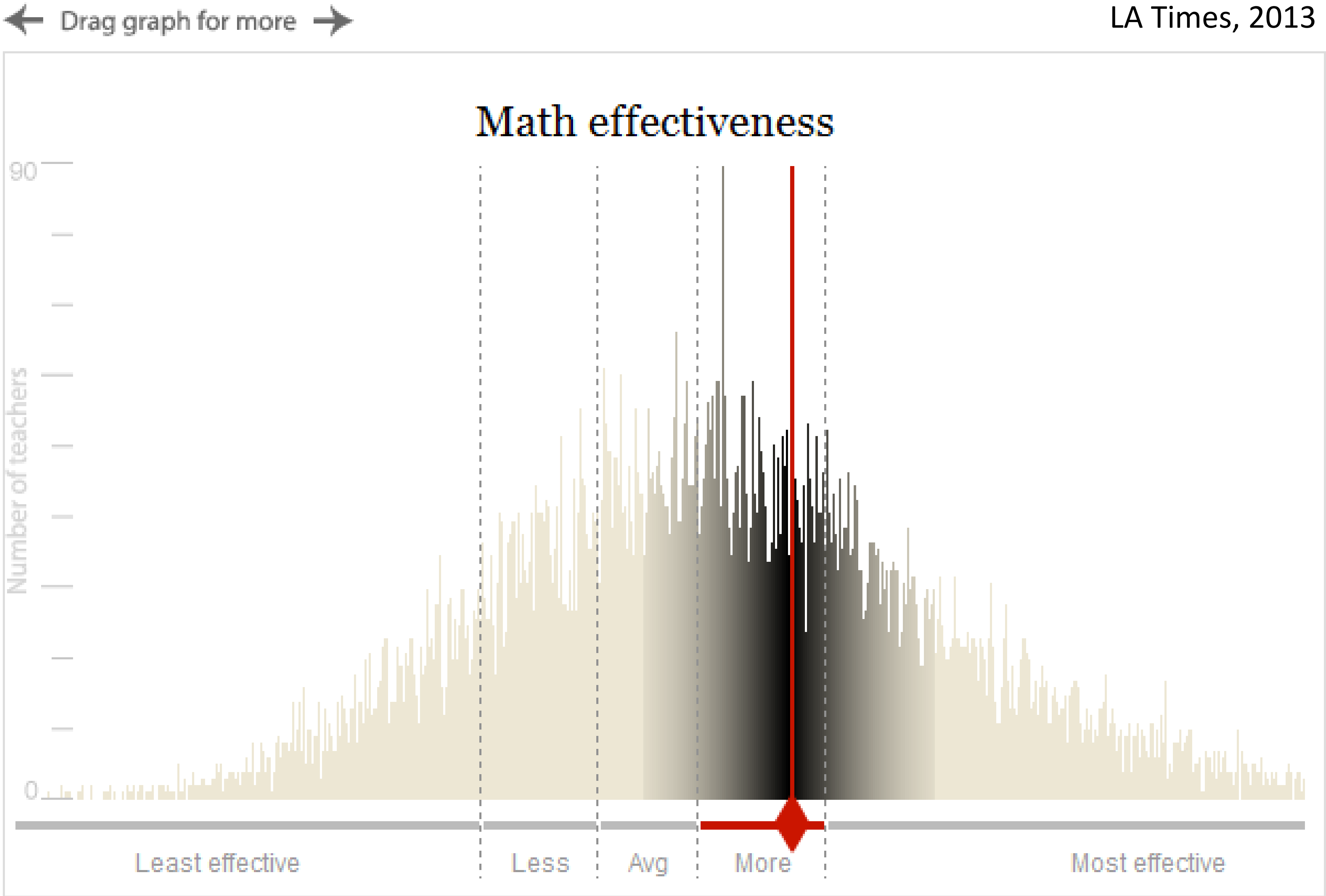
300

299 *Proficient*

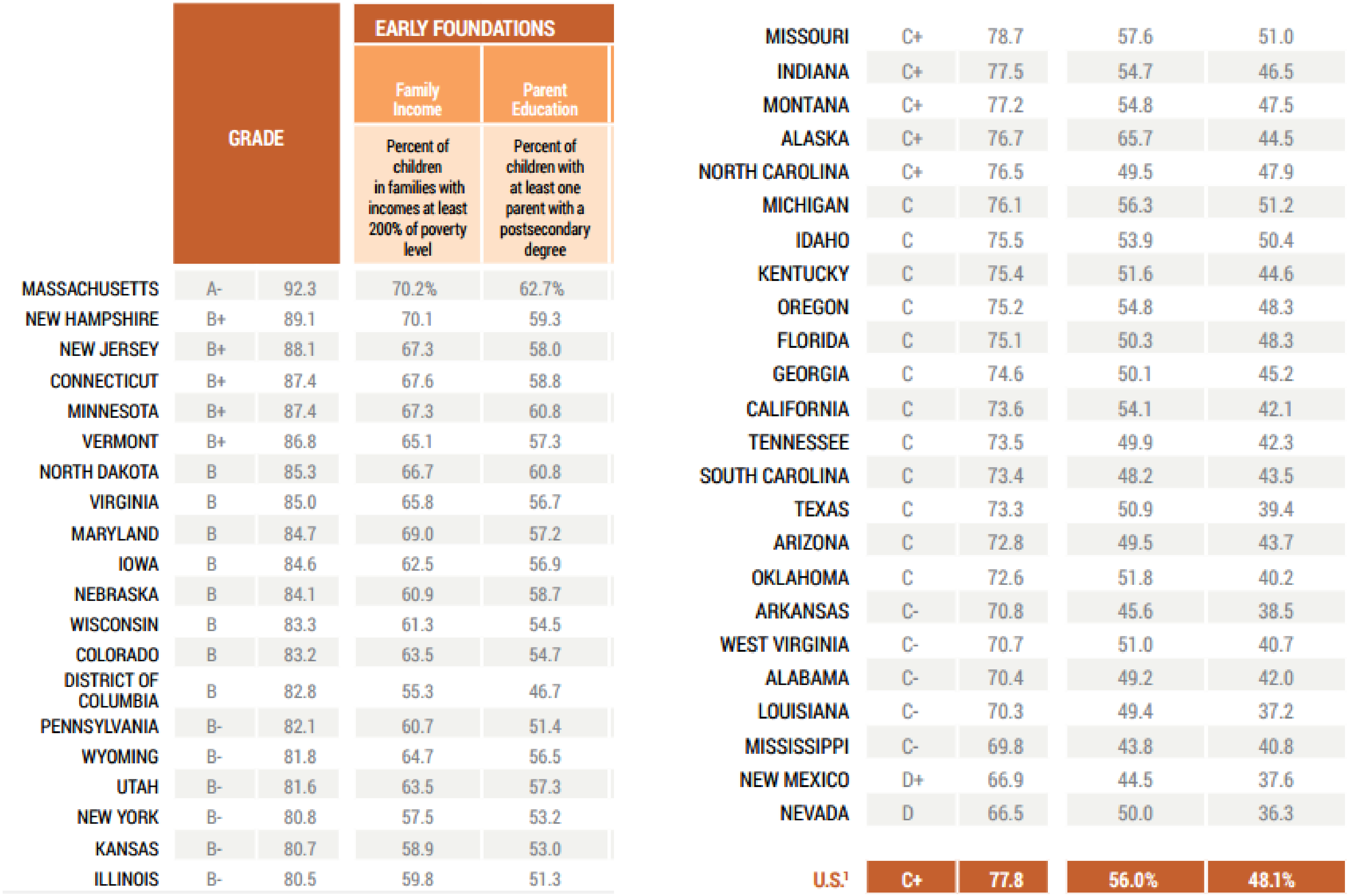
NAEP Grade 8 Item Map: http://nces.ed.gov/nationsreportcard/itemmaps/

9) Anchor scales with norms and criteria

LA Times, 2013



9) Anchor scales with norms and criteria



	GRADE		EARLY FOUNDATIONS	
			Family Income	Parent Education
			Percent of children in families with incomes at least 200% of poverty level	Percent of children with at least one parent with a postsecondary degree
MASSACHUSETTS	A-	92.3	70.2%	62.7%
NEW HAMPSHIRE	B+	89.1	70.1	59.3
NEW JERSEY	B+	88.1	67.3	58.0
CONNECTICUT	B+	87.4	67.6	58.8
MINNESOTA	B+	87.4	67.3	60.8
VERMONT	B+	86.8	65.1	57.3
NORTH DAKOTA	B	85.3	66.7	60.8
VIRGINIA	B	85.0	65.8	56.7
MARYLAND	B	84.7	69.0	57.2
IOWA	B	84.6	62.5	56.9
NEBRASKA	B	84.1	60.9	58.7
WISCONSIN	B	83.3	61.3	54.5
COLORADO	B	83.2	63.5	54.7
DISTRICT OF COLUMBIA	B	82.8	55.3	46.7
PENNSYLVANIA	B-	82.1	60.7	51.4
WYOMING	B-	81.8	64.7	56.5
UTAH	B-	81.6	63.5	57.3
NEW YORK	B-	80.8	57.5	53.2
KANSAS	B-	80.7	58.9	53.0
ILLINOIS	B-	80.5	59.8	51.3
MISSOURI	C+	78.7	57.6	51.0
INDIANA	C+	77.5	54.7	46.5
MONTANA	C+	77.2	54.8	47.5
ALASKA	C+	76.7	65.7	44.5
NORTH CAROLINA	C+	76.5	49.5	47.9
MICHIGAN	C	76.1	56.3	51.2
IDAHO	C	75.5	53.9	50.4
KENTUCKY	C	75.4	51.6	44.6
OREGON	C	75.2	54.8	48.3
FLORIDA	C	75.1	50.3	48.3
GEORGIA	C	74.6	50.1	45.2
CALIFORNIA	C	73.6	54.1	42.1
TENNESSEE	C	73.5	49.9	42.3
SOUTH CAROLINA	C	73.4	48.2	43.5
TEXAS	C	73.3	50.9	39.4
ARIZONA	C	72.8	49.5	43.7
OKLAHOMA	C	72.6	51.8	40.2
ARKANSAS	C-	70.8	45.6	38.5
WEST VIRGINIA	C-	70.7	51.0	40.7
ALABAMA	C-	70.4	49.2	42.0
LOUISIANA	C-	70.3	49.4	37.2
MISSISSIPPI	C-	69.8	43.8	40.8
NEW MEXICO	D+	66.9	44.5	37.6
NEVADA	D	66.5	50.0	36.3
U.S.[1]	C+	77.8	56.0%	48.1%

edweek.org

9) Anchor scales with norms and criteria (Ho)

Policy tools

- Scale anchoring and clear reporting
- Dashboards and multiple measures
- Progress and growth over status
- Lower stakes

10) Increase research capacity (Ho)

Legislation of complex, poorly understood systems is best done by enabling flexibility and responsiveness to empirical findings. Invest in research.

10) Increase research capacity

Example (National Research Council, 2011):

- "The modest and variable benefits shown by test-based incentive programs to date suggest that such programs should be used with caution and that substantial further research is required to understand how they can be used successfully."

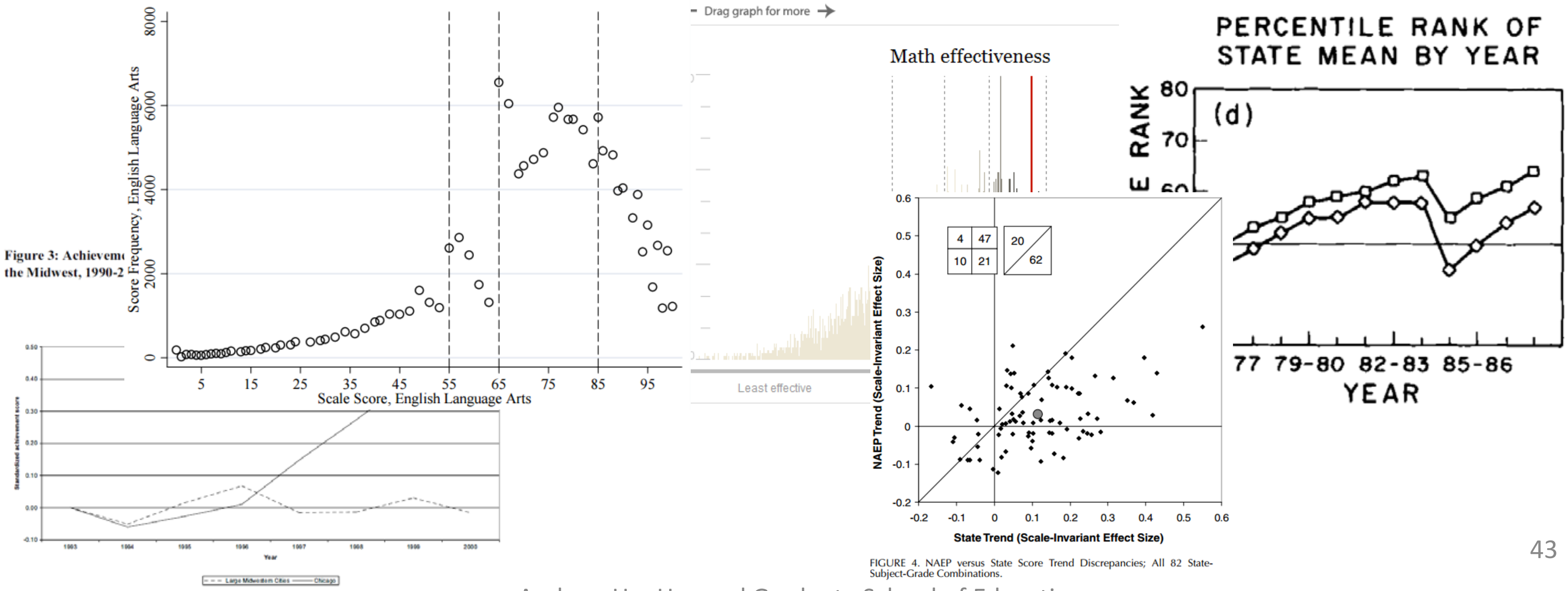


Figure 3: Achieveme
the Midwest, 1990-2

FIGURE 4. NAEP versus State Score Trend Discrepancies; All 82 State-Subject-Grade Combinations.

10) Increase research capacity

Policy tools

- Research “labs,” internal and external
- Partnerships with independent evaluators
- Nurture research relationships with other states; learn from peers.
- Longitudinal data systems

10 principles for test-based accountability systems

1. Encourage inclusion.
2. Refresh assessments yearly.
3. Use multiple measures.
4. Emphasize school improvement; downplay school rankings.
5. Emphasize student growth; also emphasize student proficiency.
6. Factor score precision into high-stakes decisions.
7. Budget for responses to unintended consequences.
8. Answer the question, “So what can I do about it?”
9. Anchor scales: What does a “B” or a “50” mean?
10. Increase research capacity.

Principles 1-7 adapted from Linn (2001)