Making Sense of and Moving Beyond Test-Based Accountability

Audrey Amrein-Beardsley, Professor

Mary Lou Fulton Teachers College, ASU

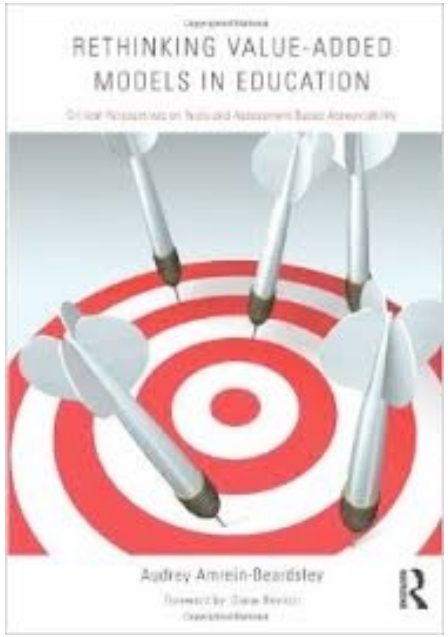
Schlechty Center

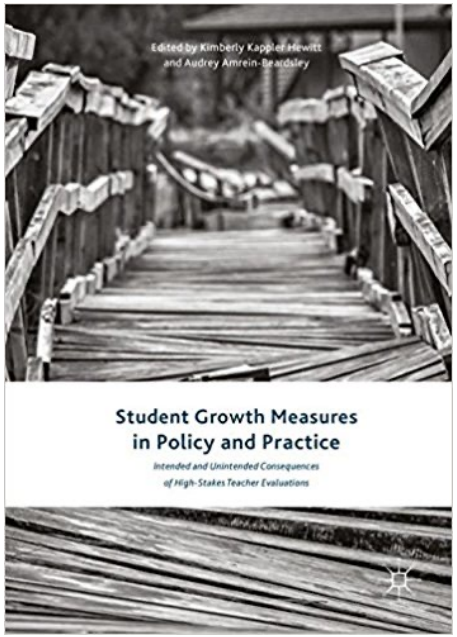
Superintendent Leadership Institute (SLN)

July 10, 2024 -- The Breakers, Palm Beach, Florida

Who I Am

- Research: Educational Tests/Measurements & Test/Measurement-Based Accountability Policies, Systems, and Practices
- Teaching: Advanced/Applied Statistics and Evaluation/Program Evaluation
- Service: Help States/Districts (and Others) Better Understand to Make Better (and More Defensible) Test/Measurement-Based Decisions





Educational Accountability History

- Sputnik (1957)
  - Preoccupation with tests as definitive indicators of educational effectiveness
- Florida (1979)
  - Minimum competency era and basic skills tests
- *A Nation at Risk* (1983)
  - Problem: U.S. public education system *not* performing, nation's global superiority "at risk," and U.S. educators failing students
  - Solutions: Increased standards and new and improved tests to hold students and educators accountable for meeting higher standards

History Cont.

No Child Left Behind (NCLB, 2001) -- Phase I

- High-stakes testing history in Texas
- Ross Perot and Former Texas Governor then U.S. President George W. Bush
- *The Myth of the Texas Miracle* (Haney, 2000)
- National calls for higher standards to which *students* were to be held accountable in grades 3-8 and once in high school in core subject areas
- Measures of Adequate Yearly Progress (AYP)
- By 2014 all *students* across all states reach 100% proficiency"
- Counterarguments
  - The *New York Times* (2002) – history predicts not intended, but unintended effects
  - The *Los Angeles Times* (2006) reveals how the Bush family "greatly profits"
  - Diane Ravitch (an NCLB author) denounces NCLB

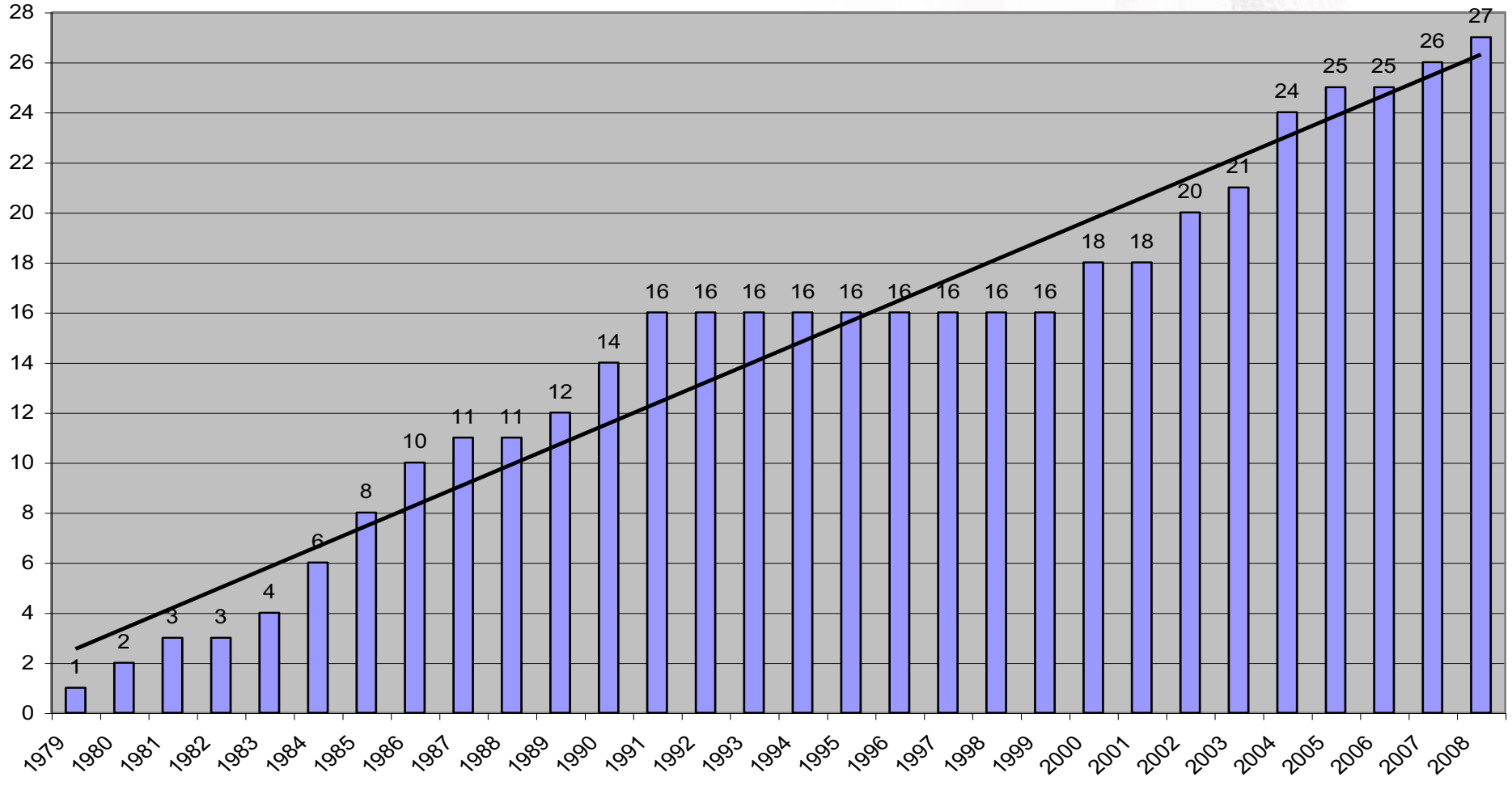
(Heroic) Arguments in Support

- Students need to be held accountable to motivate them to work harder, and learn more or teach better, respectively
- Scoring well on high-stakes tests will lead students to feelings of success, while doing poorly on such tests will lead to increased effort to work harder to obtain success
- All of this will (and should) matter more for students in U.S. schools with the highest needs and the lowest relative test performance
- High-stakes tests provide a "level playing field," yielding for all equal opportunities for all students to demonstrate their achievement
- Etc.

Low/High-Stakes Consequences

- Test scores published in newspapers
- Annual district report cards (e.g., A-F Grading System, Florida/Jeb Bush)
- High school graduation exams
- Grade-to-grade promotion exams (e.g., grades 3, 5, 8)
- Open enrollment policies (i.e., rights to move from “failing” schools)
- Teacher and administrator merit/bonus pay (cuts and bonuses)
- Teacher tenure revoked
- School accreditation revoked
- School probations, reconstitutions, and closures
- Teacher and administrator terminations (based on AYP)
- Etc.

Number of States With High School Graduation Exams
1979-2008



By-State Reactions to NCLB

States in which policymakers embraced NCLB tenets:

- Mainly in the southwest/east (i.e., Sunbelt states, not including California, but also including New York, Tennessee, Washington DC, etc.)
- Allocate less money than the national average per pupil
- Pay teachers less salary than the national average
- Have more politically conservative governments
- Have higher populations and levels of population growth
- Have higher percentages of African American, Latino/a, and American Indian students
- Have higher percentages of students living in poverty
- Have higher percentages of immigrant student populations
- Have higher percentages of students for whom English is a second language
- Have lower levels of academic achievement (still)

(Un)Intended Consequences

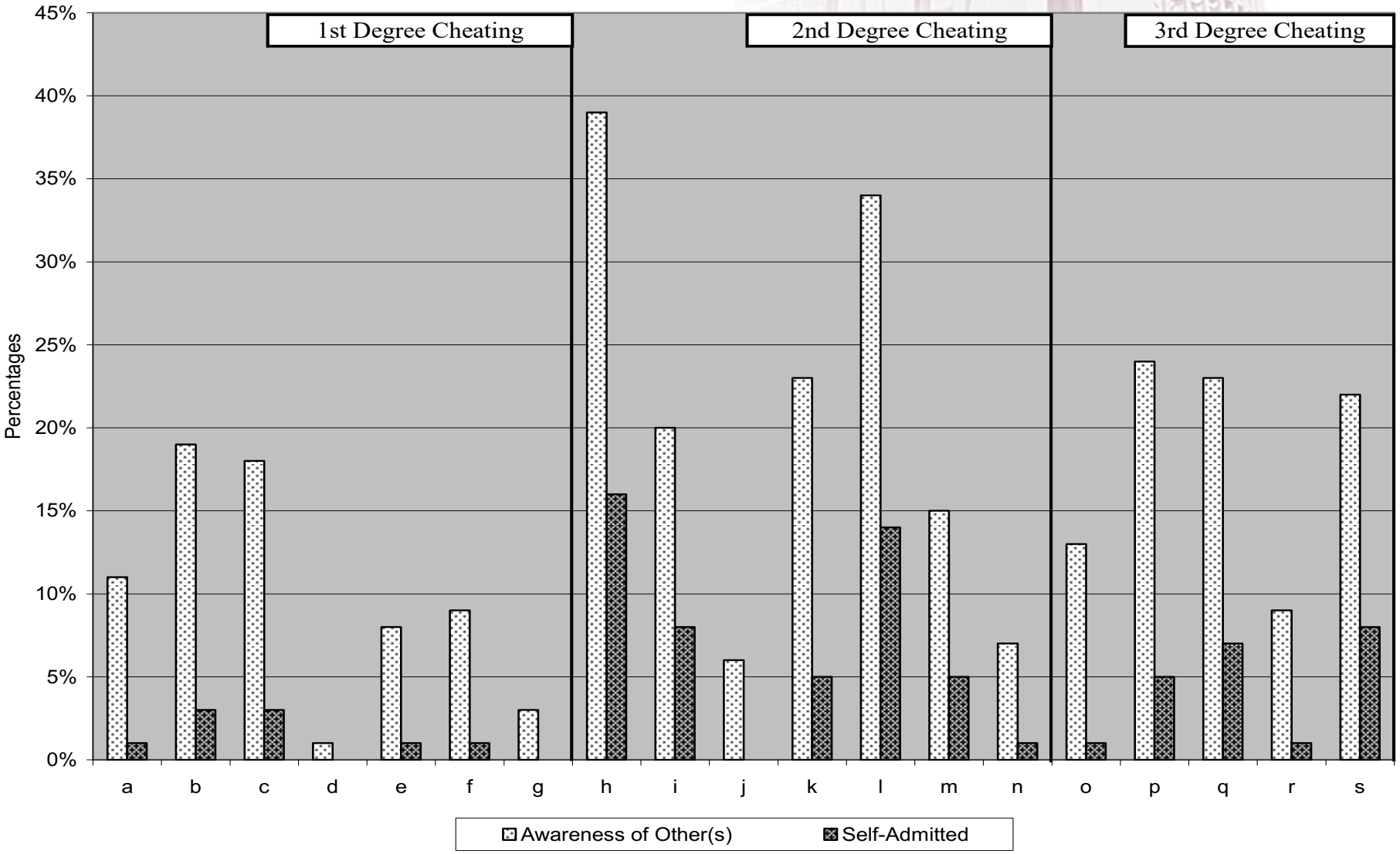
Intended

- Improved student learning and achievement
- Decreased achievement gap

Unintended

- Cheating
- Teaching to the Test
- Narrowing of the Curriculum
- Student Exclusion/Expulsion
- Student Attrition, Dropout, and GED Enrollment
- Anxiety
- Teachers Leaving Teaching, or Certain Subjects or Grade Levels

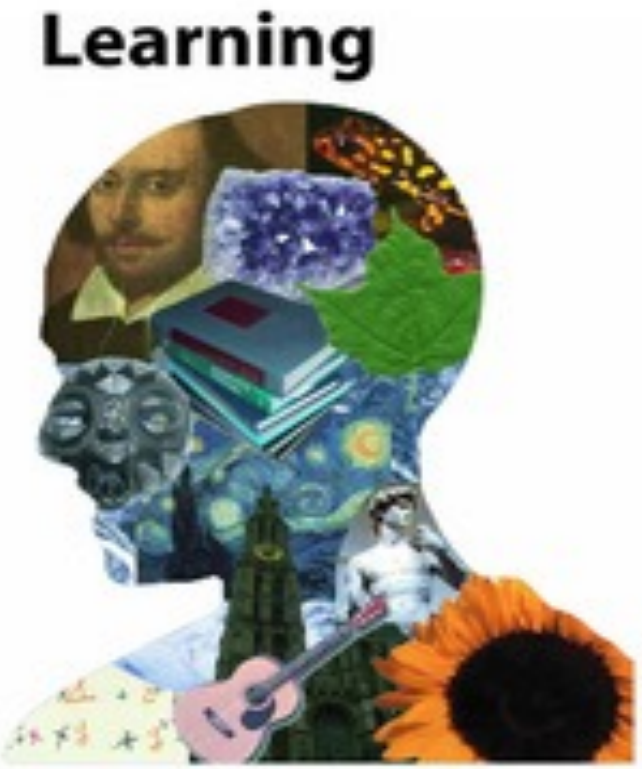
Atlanta's 2005 Cheating Scandal

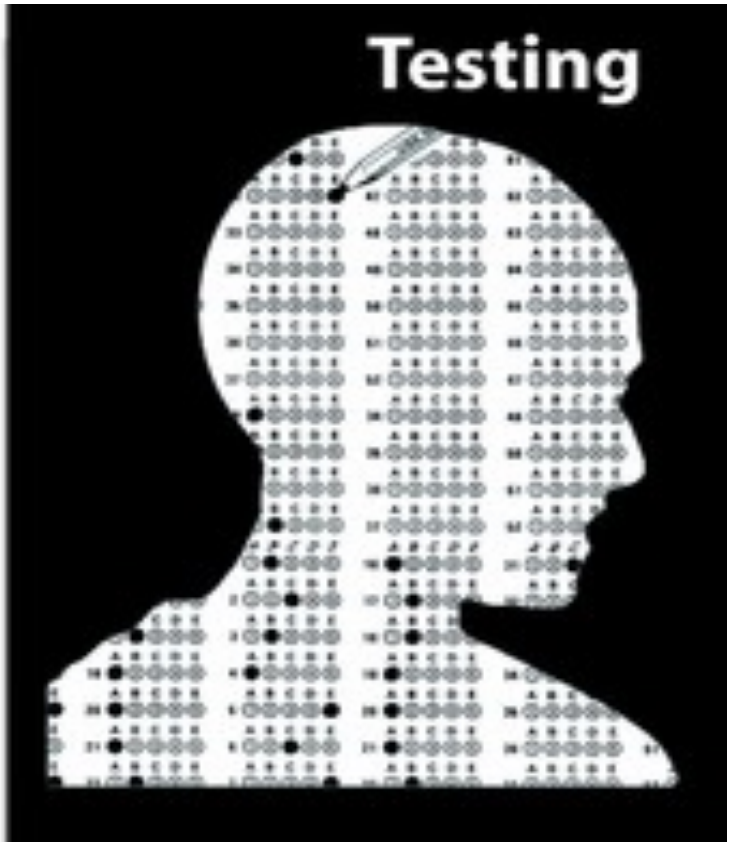

1st Degree Cheating
2nd Degree Cheating
3rd Degree Cheating
Percentages
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
Awareness of Other(s)
Self-Admitted

Teaching to the Test (not to be confused with teaching to the standards)

Narrowing of the Curriculum

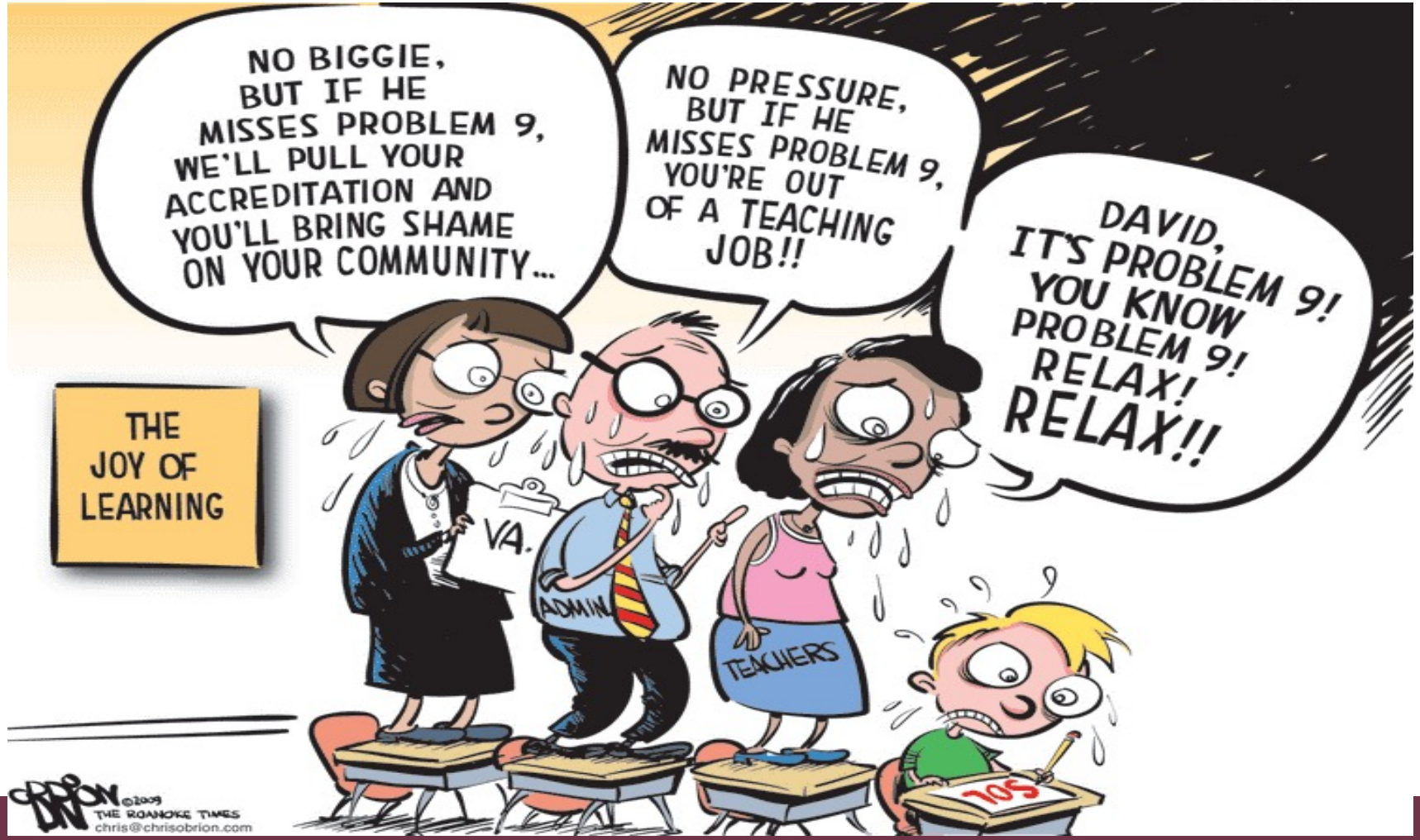
- Across subject areas (e.g., social studies, science, art, music, foreign language, physical education, recess)
- Within content-areas (e.g., performance objectives that *can be* tested)



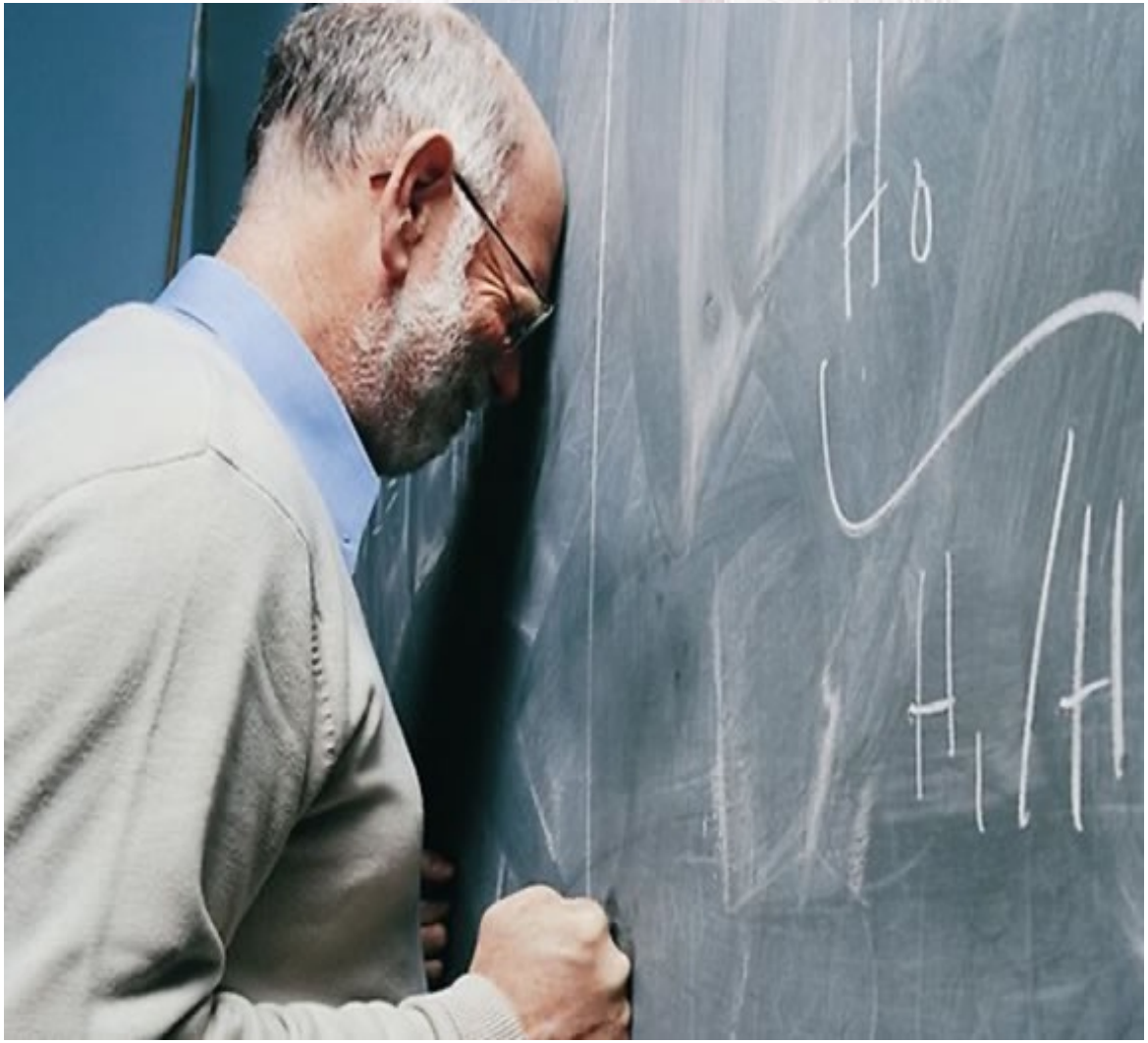


Increased Student Attrition, Student Dropout, and GED Enrollment Rates

Test Anxiety

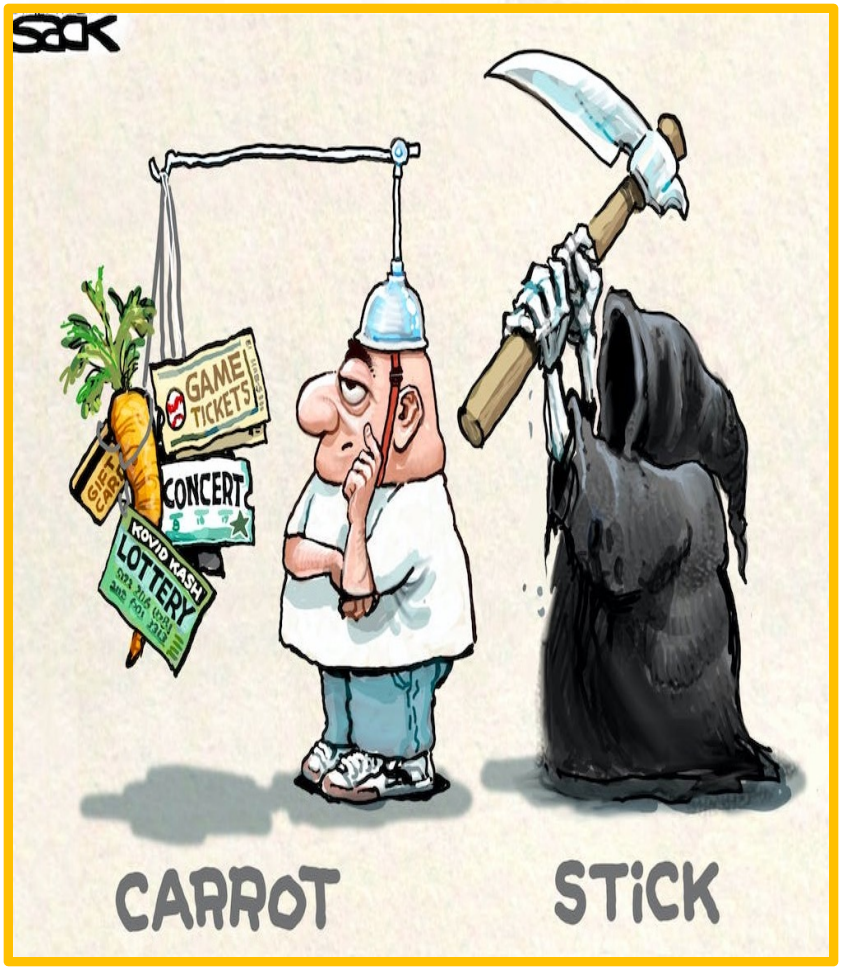


"Good" Teachers Leaving Subjects (where tests count), Grade Levels (where tests count), or Teaching Altogether (because tests count "too much")



Theory of Change

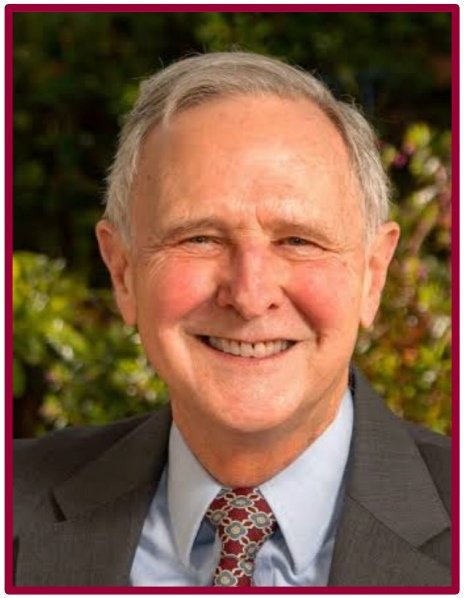
By holding states, districts, schools, administrators, teachers, and students accountable for meeting perpetually higher standards, as measured by student performance on perpetually improved tests, especially with high-stakes attached to motivate increased achievement, states will take their students' learning more seriously, administrators will lead and supervise their school personnel better, teachers will teach their students better, and students will take their learning more seriously, after which students will achieve more, after which the U.S. will reclaim its global standing among industrialized nations.



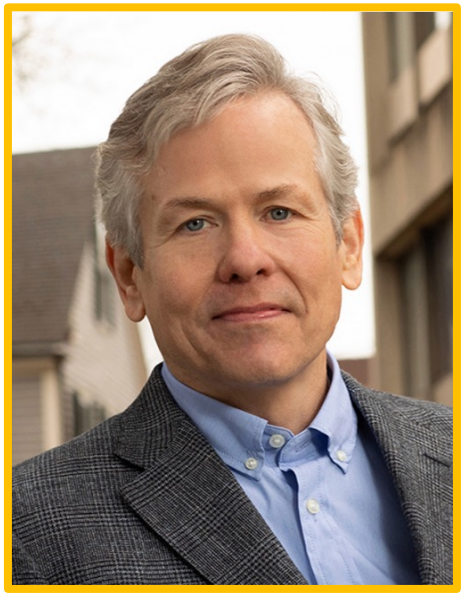
Educational Accountability "Influencers"

Policies and policy rhetoric primarily informed by the scholarly, political, & lobbying efforts of primarily *econometricians* conducting research in educational policy

Raj Chetty, Harvard



Eric Hanushek, Stanford



Thomas Kane, Harvard

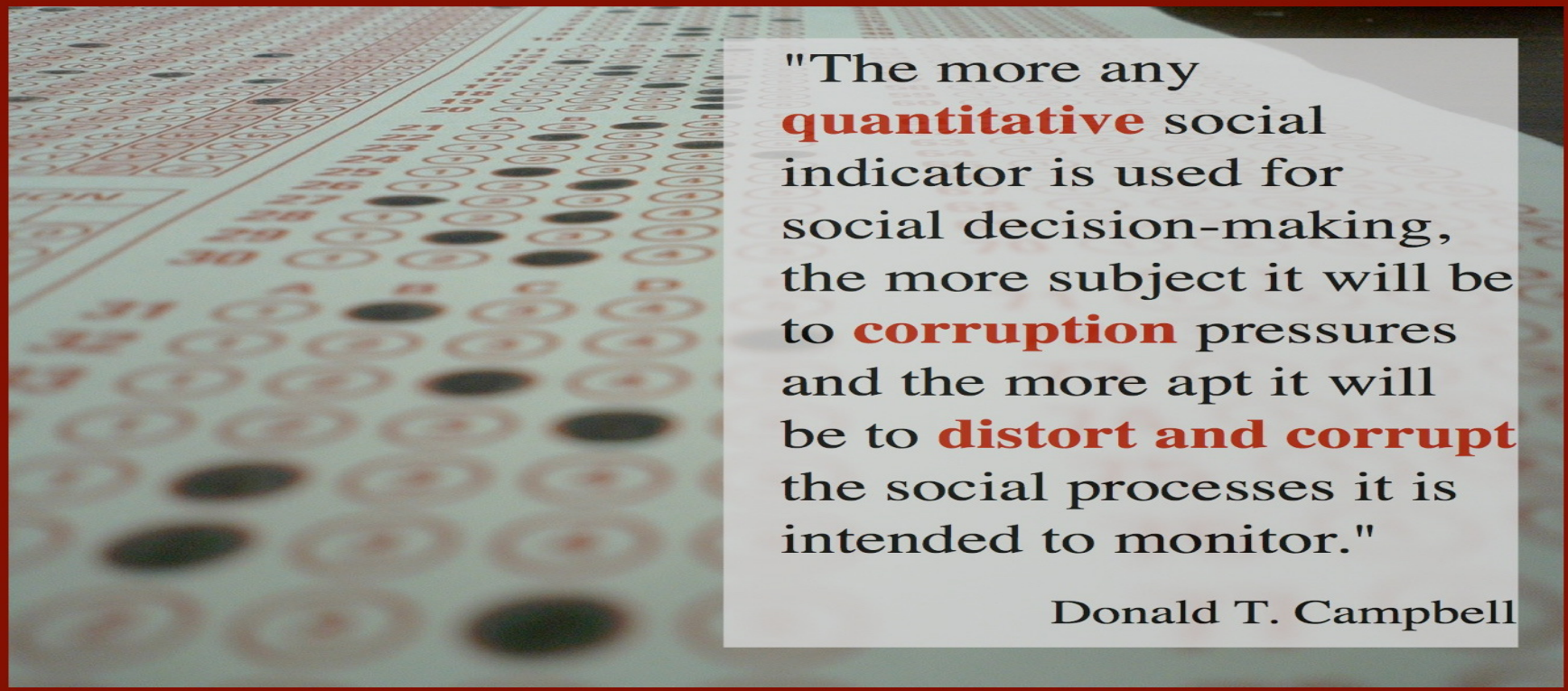
William Sanders, SAS Analytics

Educational Accountability "Contrarians"

Amrein-Beardsley, A., Ryan Lavery, M., Holloway, J., Pivovarova, M., & L. Hahs-Vaughn, D. (2023).
Evaluating the validity evidence surrounding use of value-added models to evaluate teachers: A systematic review. *Education Policy Analysis Archives.*

ASU Arizona State University

Campbell's Law



Accountability History Continued

NCLB "Fails," Enter Race to the Top (2011) -- Phase II

- NCLB's AYP measures fail
- AYP to be replaced by growth or value-added models (VAMs)
- National calls for higher standards to which *teachers* were to be held accountable in grades 3-8 and once in high school in core subject areas
- Funded states had to promise to use test scores for *teacher-level* consequences (e.g., teacher evaluation, termination, and compensation)
  - $4.35 billion in taxpayer revenues
- States also "excused" from *states' failures* to meet NCLB's prior "100 percent proficiency" goal-by-2014 via federal waivers

What is a Value-Added Model (VAM)?

- A statistical procedure to estimate a teacher's effectiveness based on "growth" in large-scale student-level test scores
  - Effective teaching or "value-added" = test score increases beyond predictions
  - Ineffective teaching, or negative/detracted "value-added" = test score increases below predictions
- Prior achievement always used to make predictions, *sometimes* along with other student demographic variables (e.g., SES, ELL, SPED)
  - Classroom/School level variable sometimes included (e.g., nested effects)

(Heroic) Assumptions in Support

- High-stakes tests are good measures of that which teachers teach
- Measuring growth using high-stakes tests are good measures of teachers' *causal* impacts on students' learning and achievement
- Teachers can use growth outputs to improve themselves and provide better instruction for students
- VAM-based data will increasingly help teachers to increasingly grow themselves, and cause greater student learning and achievement as a result
- Repeat…

Key Issues

Reliability

- Teachers classified as adding value have 25-60% chance of being classified opposite thereafter.
  - Teachers earning merit pay compare receiving monies to "winning the lottery."
  - Teachers profess not doing anything differently as their VAM-based rankings "jump around."

Validity

- Data Contradictions: Low correlations with other teacher observational/survey instrument data.

Bias

- VAM output disproportionally biased against teachers who (oft-choose to) teach gifted students, ELLs, SPEDs, students who receive free or reduced lunches, students retained in grade, etc.

Fairness

- Inability to assess teachers of non-core, non-tested subject areas and in ineligible grades
- 70% of all teachers ineligible

Transparency

- Difficult to access, interpret, and understand (e.g., "vague," "confusing," and "unclear")
- Principals deferred to "for help," but also reportedly unable

Formative Use

- Not one state able to articulate formative/professional development policies or plans (2014)
- About 60% of Houston teachers in, using "the best" VAM, report not using their data whatsoever

Reliability

	2006-2007	2007-2008	2008-2009	2009-2010
	Grade 5	Grade 4	Grade 3	Grade 3
Reading	-1.15	-0.96*	+2.03	+1.81
Math	-2.03	+0.68*	+0.16*	+3.46
Language Arts	+1.12	-0.49*	-1.77	-0.20*
Science	+2.37	-3.45	n/a	n/a
Social Studies	+0.91*	-2.39	n/a	n/a

- Teacher A was a teacher for more than 10 years, teaching elementary school in HISD since 2000.

- Teacher A added value to her students' learning 50% of the time (8/16 of EVAAS observations) and detracted value the other 50% of the time (8/16 of EVAAS observations). During Teacher A's most recent years of activity, she added more value than years prior.

- Until 2010-11, she "exceeded expectations" across every domain in terms of her supervisor evaluations. She was also given a "Teacher of the Month" award in 2010 and a "Teacher of the Year" two years prior.

Validity – (Mis)Attributions

- About 86%—99% percent variation in test scores is from outside-of-school factors.
- About 1%—14% of the variation in test scores is attributable to school-level factors, including but not limited to teachers.

American Statistical Association, 2014

Outside-of-School Factors (Very) Strongly Correlated with Large-Scaled Standardized Test Scores:

- Family income
- Average community income
- Food insecurity
- Percent of single parents/guardians
- Percent of parents/guardians without a high school degree
- Percent of low birth-weight children in the neighborhood
- Language spoken at home
- Access to early education
- Inadequate medical, dental, vision, and mental health care
- Rates of violence & drug use in the neighborhood
- Eviction and mobility rates of families in the neighborhood
- Transportation systems to access jobs
- Etc.

Marketing Claims: Where is the Evidence?

The Education Value-Added Assessment System (EVAAS) will help to:

- "Enhance working conditions"
  - 75.5% disagreement
- "Validly identify and help remove ineffective teachers"
  - 72.9% disagreement
- "Enhance the school environment"
  - 72.5% disagreement
- "Identify excellence in teaching or leadership"
  - 68.3% disagreement

- "Increase college graduation rates by 2020"

(Un)Intended Consequences in Court

- *Education Week* (2015): 15 state-level lawsuits
  - *Teacher evaluation systems heads to the courts*
- Florida, New Mexico, New York, Tennessee, Texas, and the Houston Independent School District (HISD)
  - 14th Amendment Ruling in HISD
  - Tennessee
    - 4th & 8th teachers significantly better than 5th, 6th, and 7th teachers
    - Mathematics teachers significantly better than English/language arts teachers
    - Artificial conflation (and inflation/deflation)
- Alabama, Georgia, Nevada, Oklahoma, Pennsylvania, and likely others changed legislation, in ways to avoid potential lawsuits

History Cont.

Every Student Succeeds Act (ESSA, 2016)

- ESSA was to:
  - Reset] “Washington’s relationship with the nation’s 100,000 public schools,” its nearly 50mil public school students, and its 3.4mil teachers
  - Reduce “the federal footprint and restore local control…empowering parents and education leaders to hold schools [and teachers] accountable”
  - Replace the current national accountability system with state-led accountability systems
- All states still required to test students annually in mathematics and reading in grades three through eight and once in high school (NCLB)
- The same theories of change and the same assumptions continue to hold…
- However, some progress seems to have been made
  - Fewer states mandating/endorsing VAMs
  - More states focusing on local control
  - More states embracing the formative functions of multiple-measure accountability systems

History Cont.

Enter Jeb Bush's Excel*In*Ed "Initiatives" -- Phase III

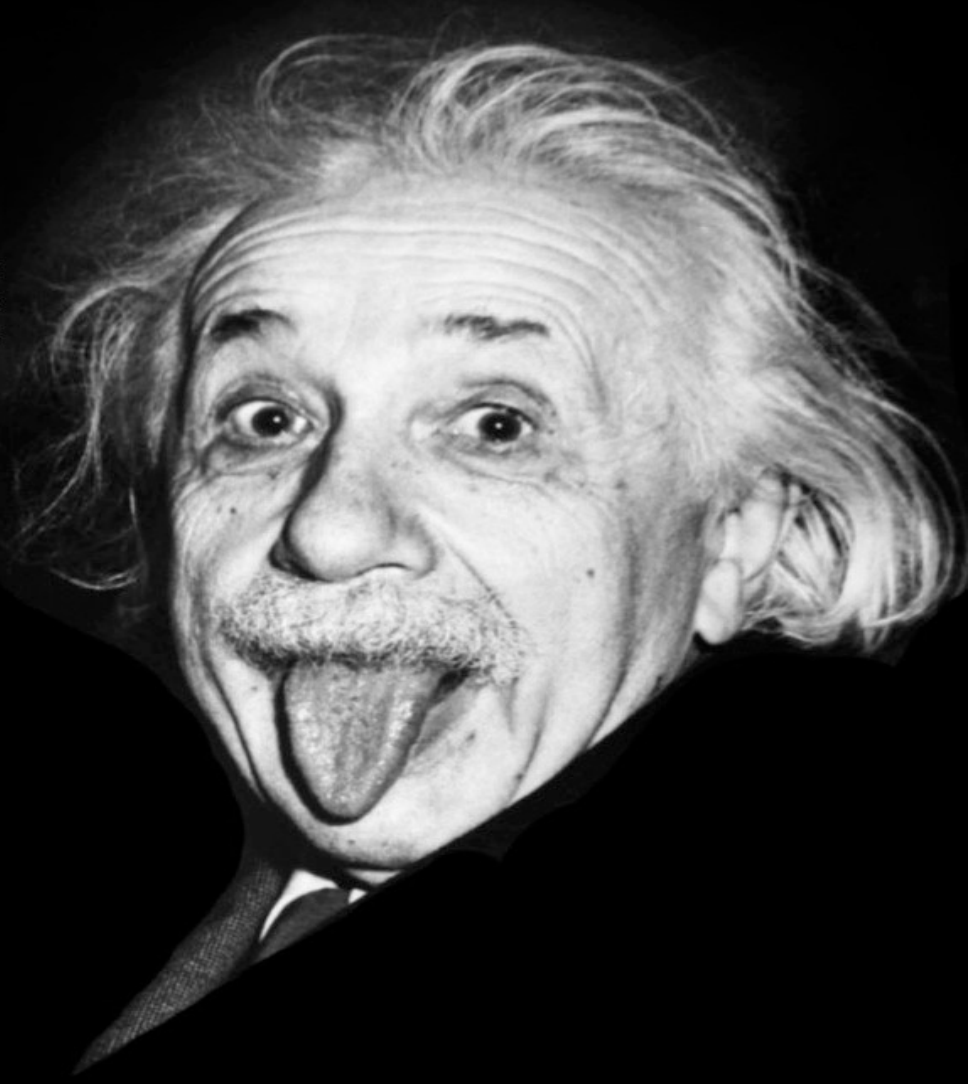
- Focus now on *school* accountability
- Pushing, state-by-state, A-F *school* letter grade accountability policies
  - $12 million per year of donations from the Bill & Melinda Gates Foundation, Michael Bloomberg Philanthropies, the Walton Family Foundation, and the Pearson, McGraw-Hill, Northwest Evaluation Association, ACT, College Board, Educational Testing Service (ETS), etc.
- Florida the first to implement an A-F *school* accountability policy in 1998
- 13 states have since adopted A-F policies
- Preliminary research evidences unimpressive results (again)
  - About half of states increase National Assessment of Educational Progress (NAEP) achievement post policy implementation
  - About half decrease on the NAEP
  - In sum, states' adoption of A-F policies yield perfectly random results

Some (Evidence-Based) Conclusions

Research on these and other similar policy reforms, perpetually studied for the past, now nearly 45ish years, yields evidence that:

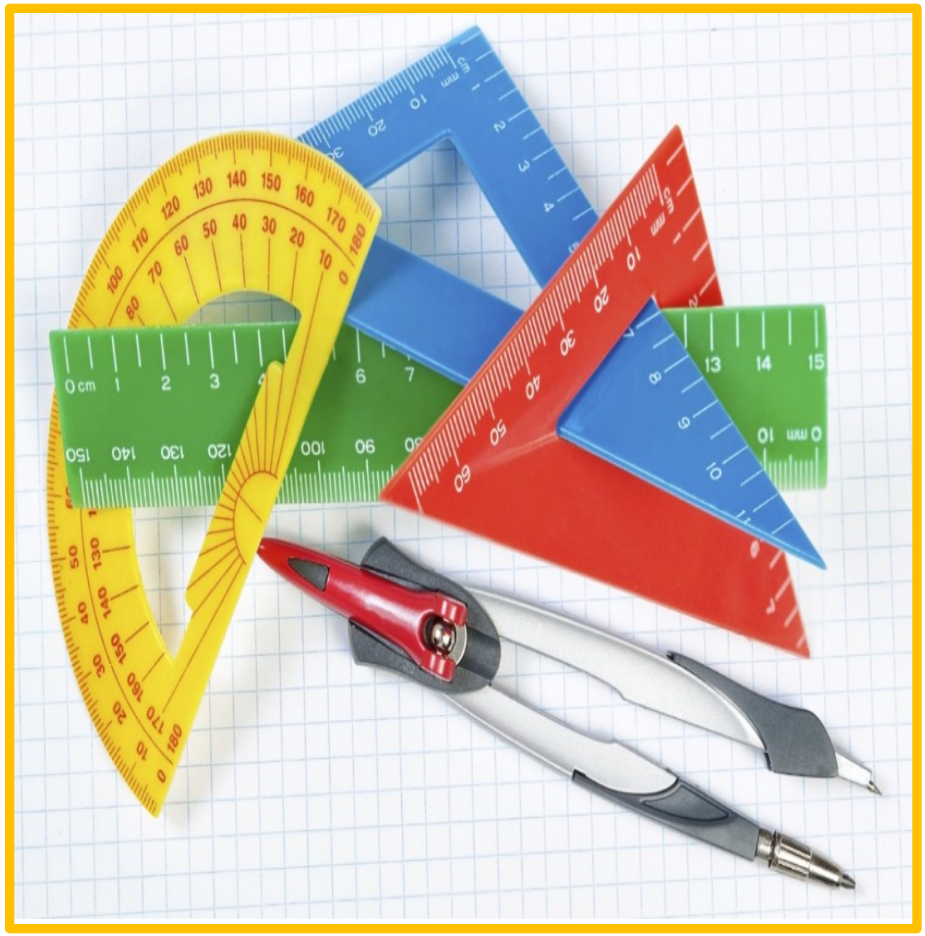
- U.S./State policymakers operate in highly ahistorical ways
- *All* aforementioned (heroic) assumptions are false, a good deal of the time
  - U.S.'s NAEP scores (i.e., flatline/decline/incline post all policy initiatives, although few do disagree)
  - U.S.'s continued/purported "subpar performance" on international tests as compared to other industrialized nations
- Researchers often evidence a lack of intended effects but, rather, a plethora of oft-adverse, unintended effects instead

INSANITY: DOING THE SAME THING OVER AND OVER AGAIN, AND EXPECTING DIFFERENT RESULTS.



Solutions?

Multiple Measures



- Using multiple measures *works* regardless of unit of measurement (e.g., student, teacher, administrator, school)
- Measures *locally* developed, implemented, tested, revised, owned, *and* democratically used for purposes of accountability
- Measures *may* include:
  - Student learning indicators
  - Observational data (admin/peer)
  - Survey-based data
  - Community-based measures
  - Leadership indicators
  - Service indicators
- Measurement weaknesses offset others
- *Never* marginalize human judgment

American Educational Research Association et al. (2014)

Advice for Educational Leaders

Critical Consumption v. Blind or Uninformed "Trust"

Caretakers of Taxpayer Revenues

Multiple Measures including Quantitative and Qualitative Indicators

External Research (also on all components)

- TAP Factor Analysis study just published here and Danielson study forthcoming

Follow the *Standards for Educational and Psychological Testing*

Local Control

Avoid Litigation (i.e., Avoid Using/Implementing Indefensible, High-Stakes Consequences Until Evidence-Based or Warranted)

Where is the Evidence?

Where are the (publicly available) technical reports, with the:

- Statistics demonstrating model reliability/consistency and what do such statistics mean, with implications for policy and practice
- Evidence regarding validity (e.g., tests can be used to measure teachers' impacts on student achievement)
- Statistics demonstrating inter-indicator correlations to evidence that the "multiple measures" within any systems are appropriately correlated
- Statistics demonstrating a lack of bias (e.g., teachers of gifted students, ELLs, special education students, certain grade levels, and certain subject areas are as likely as their colleagues to demonstrate value added).
- Etc…

No leader should endorse/buy into any accountability model without proof that it works, especially prior to the attachment of high-stakes consequences

Some Final Words on Accountability

While critical of test-based accountability programs, I do not oppose "accountability."

- A proponent of accountability, but not the current, bureaucratic version that now exists.
- Care about the transformation of schools from bureaucratic organizations into learning organizations.
  - It is the highly engaged school and classroom where profound learning occurs, and it is the well-managed school and classroom where learning is superficial.
  - In a "school as factory" model, the core business *is* simply testing and remediating.
- If we are going to hold someone accountable, we must understand the conditions under which they learn and work.
- We must avoid simple solutions to capture or explain highly complex phenomena.
  - We are dealing with variables that are highly complex:
    - Somebody teaching (variations in instruction, curriculum, etc.) something (variations in subject areas, grade levels, etc.) to somebody (variations in students).
  - To think that the complexities of such variables, and interactions among variables, can be captured by a test is absurd.
- Multiple measures, as per research-based practice, so as also to inform research-based policy

Thank you!

Questions?

Email: audrey.beardsley@asu.edu