Center for Assessment

Session 3 (Part 2)
Consequential Uses of Assessment: A Friendly Debate

Reidy Interactive Learning Series (RILS) Conference
Portsmouth, NH, September 26-27, 2024
AC Marriott Hotel

Overview (Part 2)

Discussant 20′

Table Reflections 20′

Group Share-outs 20′

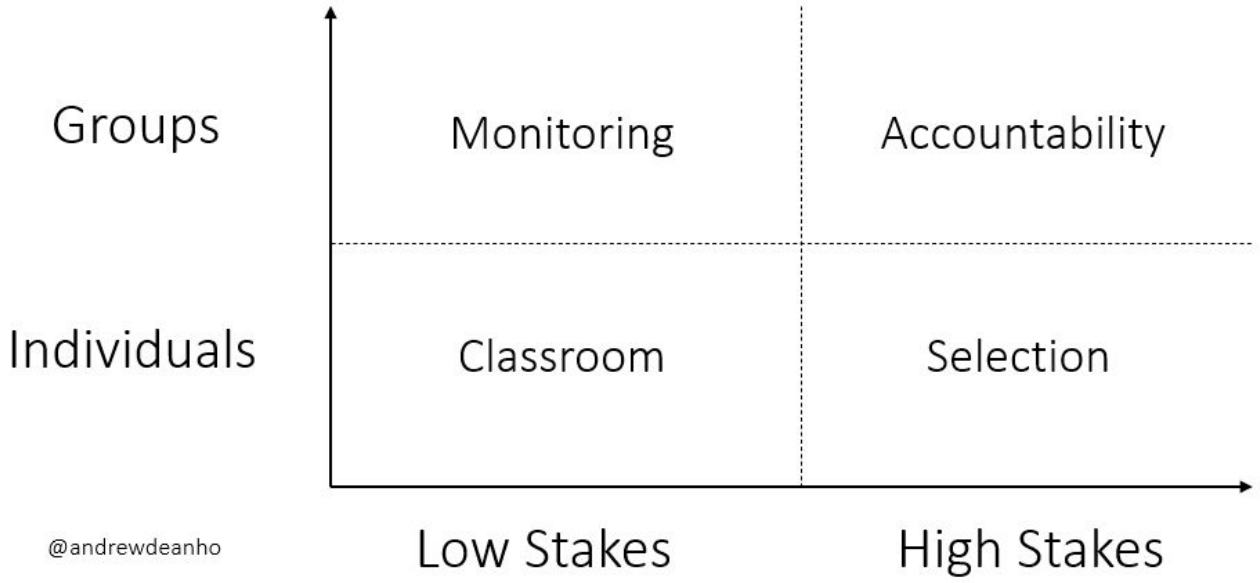
Andrew Ho
Harvard University

Guiding Questions

- What arguments *for* each proposition did you find persuasive?
- What arguments *against* each proposition did you find persuasive?
- What arguments from the discussant resonated with you?
- What questions or comments should be further discussed?

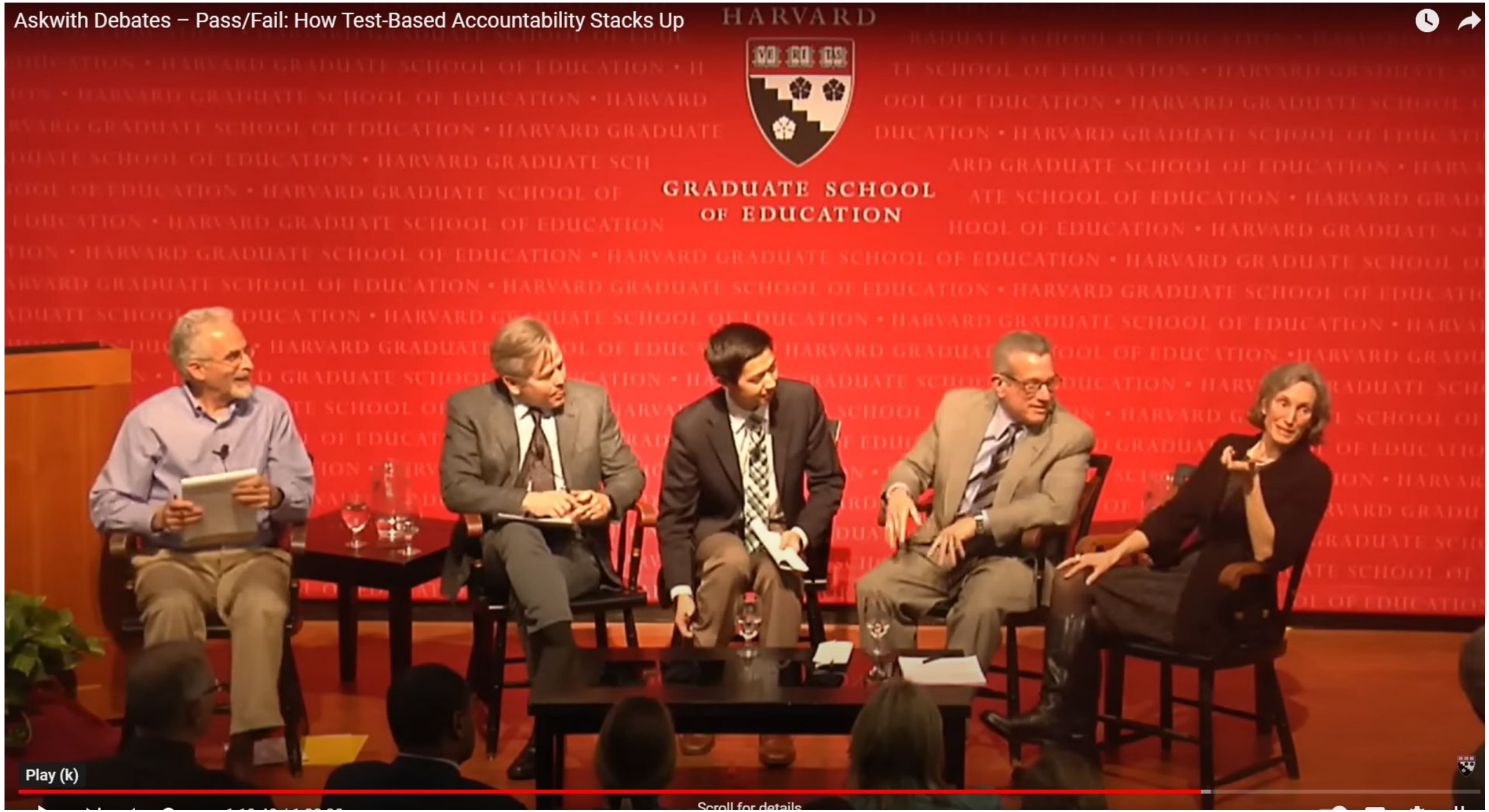
Discussant: Andrew Ho

Consequences or Responses?
Resolving a "Friendly Debate" About High-Stakes Testing

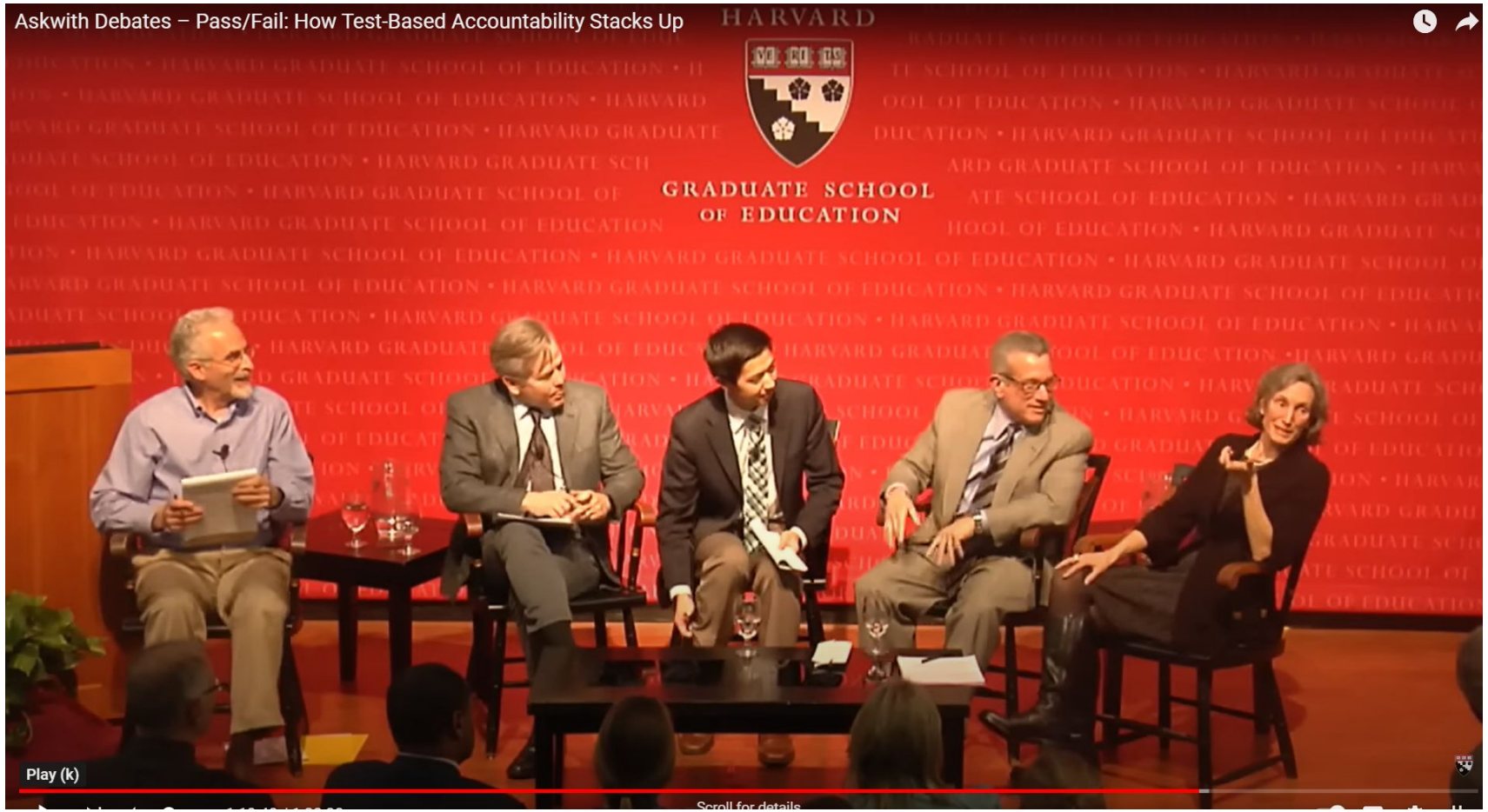


Andrew Ho, *Charles William Eliot Professor of Education* Twitter/Bluesky/LinkedIn: @andrewdeanho
Harvard Graduate School of Education
September 26, 2024
Reidy Interactive Learning Series, National Center for the Improvement of Educational Assessment

This reminds me of a time… [video link]



This reminds me of a time… [video link]



Can't we all agree on my riff on Elmore's (2004) principles for reciprocal accountability?

1. Multiple measures
2. Achievable targets
3. No stakes without support

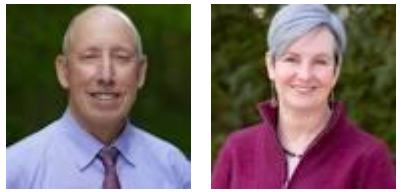
Three Ws (Ho, 2002) and Four Quadrants (bit.ly/hoquadrants)

Who uses Which scores for What purpose?

Groups (Prop. 2)

Individuals (Prop. 1)

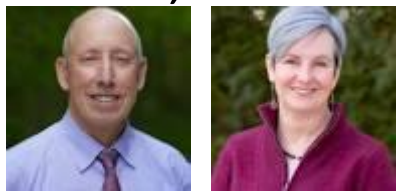
The **Monitoring** Quadrant (e.g., NAEP, TIMSS, PISA)



The **Accountability** Quadrant (e.g., teacher value added, state tests)

The **Classroom** Quadrant (e.g., formative, diagnostic, feedback)

The **Selection** Quadrant (e.g., admissions, awards, remediation, certification)

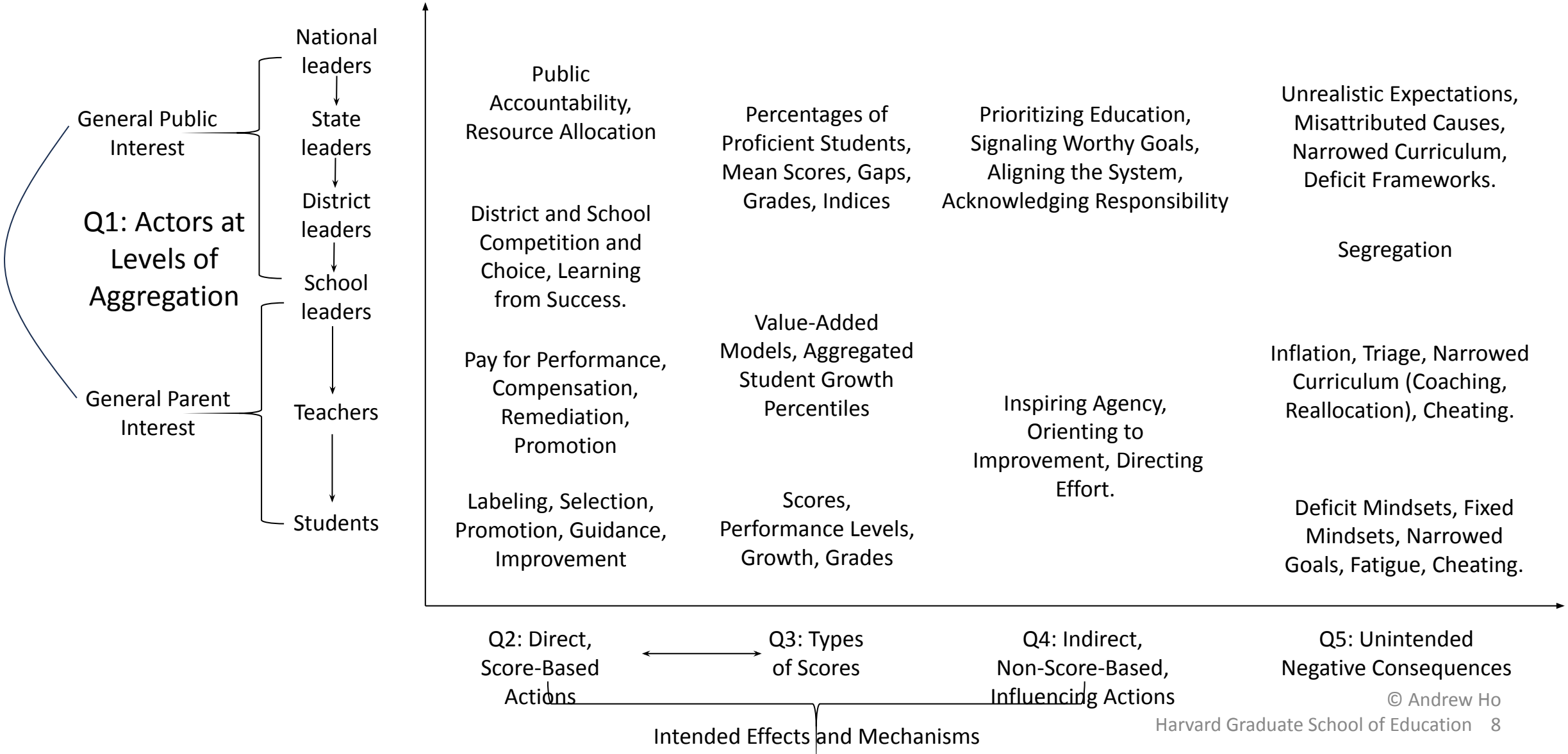


Low-Stakes

High-Stakes

A comprehensive framework for test-based accountability (EM5: Ho & Polikoff, 2025)

1) Who is holding whom accountable? 2) By what mechanism? 3) Using which scores?
4) For what purpose? 5) With what unintended negative consequences?



Thou Shalt Not Place High Stakes Upon a Single Measure (AERA/APA/NCME, 2014

Standard 12.10

In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information.

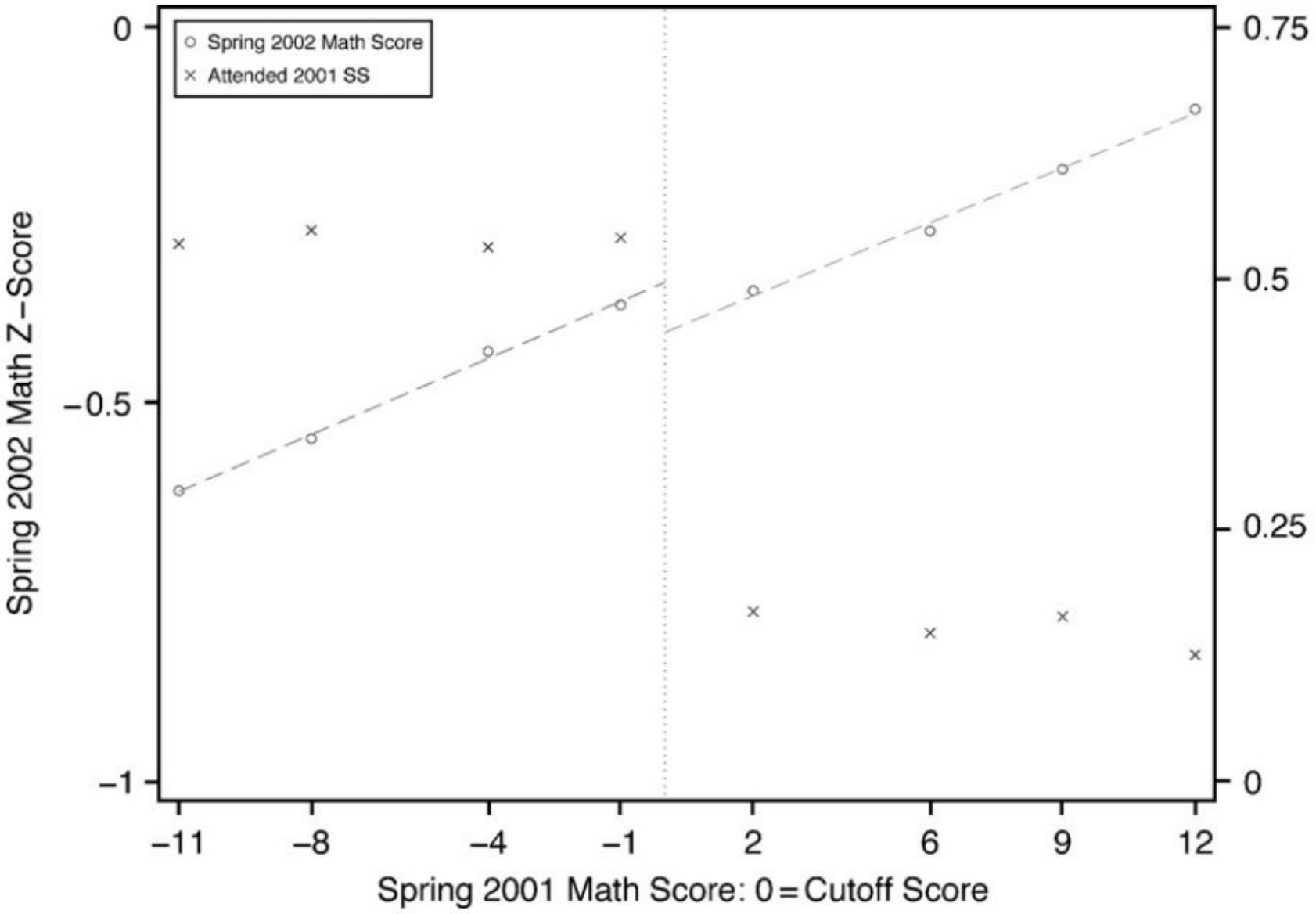
Standard 13.9

In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation.

"When tests are used for promotion and graduation, the fairness of individual score interpretations can be enhanced by

(a) providing students with multiple opportunities to demonstrate their capabilities through repeated testing with alternate forms or other construct-equivalent means;
(b) providing students with adequate notice of the skills and content to be tested, along with appropriate test preparation materials;
(c) providing students with curriculum and instruction that afford them the opportunity to learn the content and skills to be tested;
(d) providing students with equal access to disclosed test content and responses as well as any specific guidance for test taking (e.g., test-taking strategies);
(e) providing students with appropriate testing accommodations to address particular access needs; and
(f) in appropriate cases, taking into account multiple criteria rather than just a single test score. (AERA/APA/NCME, 2014, p. 187)

Can’t some high-stakes test-based policies help students? (Matsudaira, 2008)



Fraction attending summer school and achievement outcomes by 2001 math score: Grade 5—all students.

Using administrative data from a large school district, I exploit the fact that students are mandated to attend summer school based on a discontinuous function of their score on year-end exams to identify the effect of summer school attendance on achievement. I find an average effect of about .12 standard deviations for both math and reading achievement, an effect size on the low end of the range of prior estimates. These averages mask considerable heterogeneity, however, with effect size estimates ranging from just below zero to one-quarter of a standard deviation

Isn't School Turnaround Working? (Schueler et al., 2022)

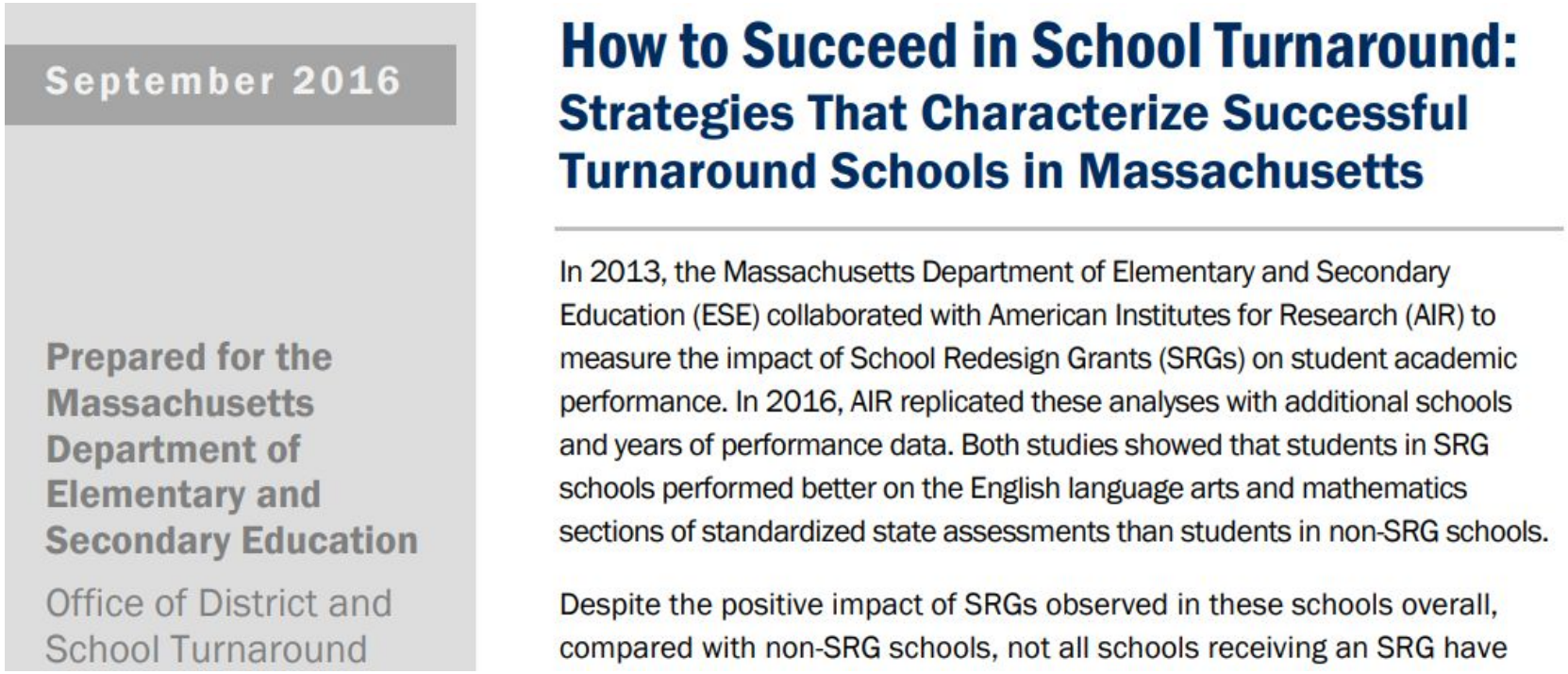
Improving Low-Performing Schools: A Meta-Analysis of Impact Evaluation Studies

Beth E. Schueler
University of Virginia
Catherine Armstrong Asher
University of Michigan
Katherine E. Larned
Harvard University
Sarah Mehrotra
Education Trust
Cynthia Pollard
Harvard University

American Educational Research Journal
October 2022, Vol. 59, No. 5, pp. 975–1010

The public narrative surrounding efforts to improve low-performing K–12 schools in the United States has been notably gloomy. But what is known empirically about whether school improvement works, which policies are most effective, which contexts respond best to intervention, and how long it takes? We meta-analyze 141 estimates from 67 studies of post–No Child Left Behind Act turnaround policies. On average, policies had moderate positiv effects on math and no effect on English Language Arts achievement on high-stakes exams. We find positive impacts on low-stakes exams and no evidence of harm on nontest outcomes. Extended learning time and teacher replacements predict greater effects. Contexts serving majority-Latina/o populations saw the largest improvements. We cannot rule out publication bias entirely but find no differences between peer-reviewed versus nonpeer-reviewed estimates.

How has it worked?



September 2016

Prepared for the Massachusetts Department of Elementary and Secondary Education

Office of District and School Turnaround

How to Succeed in School Turnaround: Strategies That Characterize Successful Turnaround Schools in Massachusetts

In 2013, the Massachusetts Department of Elementary and Secondary Education (ESE) collaborated with American Institutes for Research (AIR) to measure the impact of School Redesign Grants (SRGs) on student academic performance. In 2016, AIR replicated these analyses with additional schools and years of performance data. Both studies showed that students in SRG schools performed better on the English language arts and mathematics sections of standardized state assessments than students in non-SRG schools.

Despite the positive impact of SRGs observed in these schools overall, compared with non-SRG schools, not all schools receiving an SRG have

1) Leadership, Shared Responsibility, and Professional Collaboration
2) Intentional Practices for Improving Instruction
3) Student-Specific Supports and Instruction to All Students
4) School Climate and Culture

Are these “consequences” or “responses”?

- It is what we do for students and schools that “fail” that determines downstream outcomes.
- Is it, “this test has consequences,” or “people, policies, and systems respond to test scores.”
- So let us evaluate and improve how we respond.
- “Measurement must be qualitative, then quantitative, then qualitative again” (Ho et al., 2024).

Table Reflections & Share-outs

Guiding Questions

- What arguments *for* each proposition did you find persuasive?
- What arguments *against* each proposition did you find persuasive?
- What arguments from the discussant resonated with you?
- What questions or comments should be further discussed?

Center for Assessment

Session 3 (Part 2)
Consequential Uses of Assessment: A Friendly Debate

Reidy Interactive Learning Series (RILS) Conference
Portsmouth, NH, September 26-27, 2024
AC Marriott Hotel