Challenges and Opportunities of Observational Teaching Metrics to Assess and Improve Teaching

DREW GITOMER
GRADUATE SCHOOL OF EDUCATION, RUTGERS UNIVERSITY

JOSÉ FELIPE MARTÍNEZ
SCHOOL OF EDUCATION & INFORMATION STUDIES, UCLA

PRESENTED AT QUINT CONFERENCE:
ANALYZING TEACHING QUALITY – INSTRUMENTS, IMPROVEMENTS AND IMPLICATIONS. WHAT HAVE WE LEARNED?
JUNE 2024

We Regret Not Being in Oslo With All of You!

A Work Journey by Two Participant-Observers

- Excitement and promise
- Creation and refinement
- Exploration, frustration, concern, and persistence
- Acceptance
- Moving forward

Outline of Talk

- Background: Reliability and validity - what does the research say?
- Methodological issues
- Practical and policy challenges
- Understanding the inherent limits of the enterprise
- Future directions for classroom observation
  - Utility and formative contexts
  - Alternative conceptualizations of error and validity
  - Artificial Intelligence: A new frontier or the latest fad?

Caveat

- Our experiences and perspectives are informed, to a great extent, by the U.S. education context.
- Many countries have not adopted large-scale testing and accountability policies common in the United States. However:
  - some have adopted, have tried, or have considered/may consider; and
  - many assumptions, considerations, and challenges apply generally across research, policy, and practice contexts:
    - weak theories, under-conceptualized constructs, inconsistent evidence, methodological limitations, and practical challenges; and
    - a new, burgeoning field of research.

“The [classroom environment] data obtained in such records are . . . selective, inconsistent, and usually incomparable with other records. This is due to the tremendous complexity of any social behavior act and the consequent recording of different elements of these complex acts at different times."

-- Dorothy Swaine Thomas, 1929

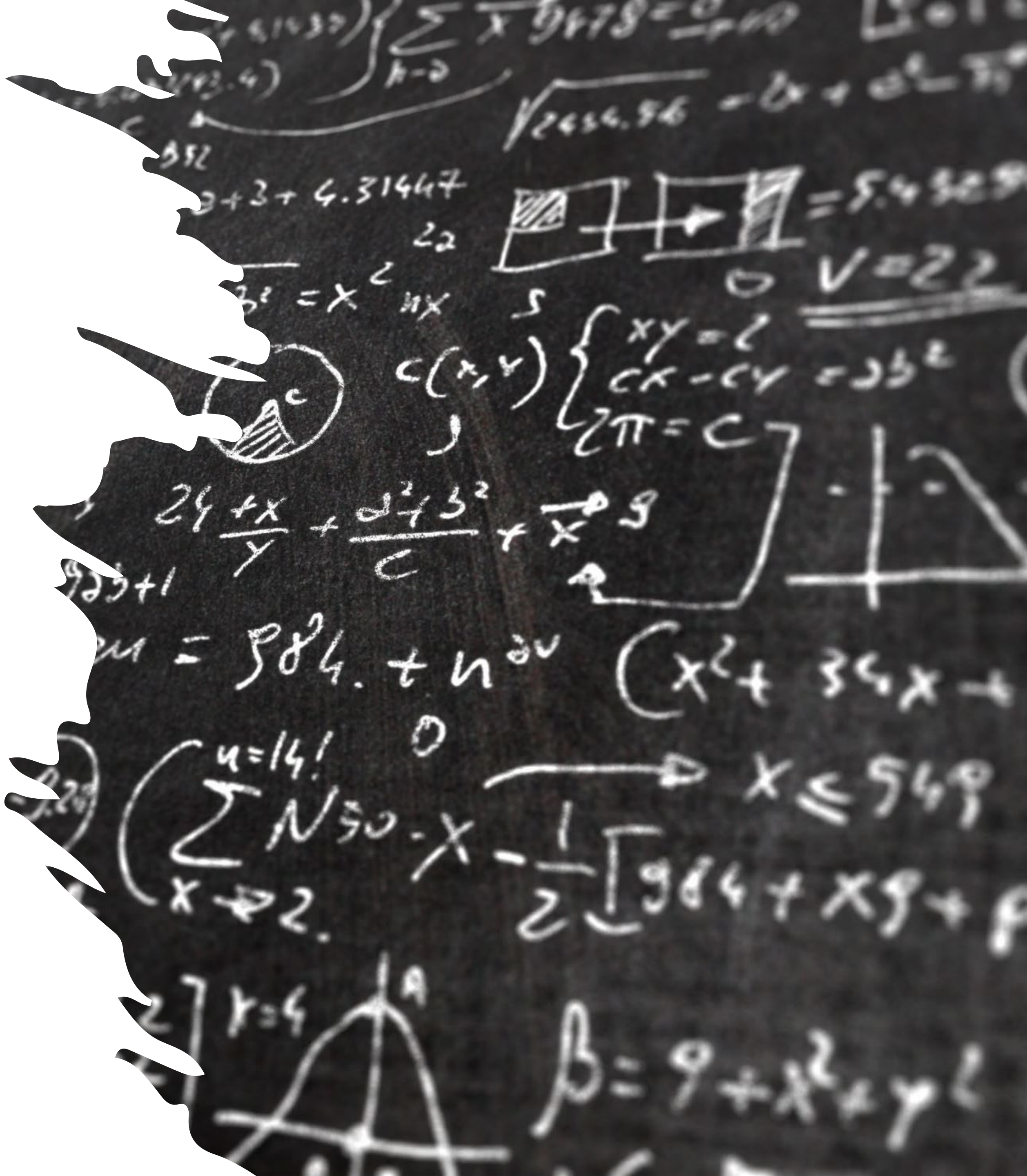
A Persistent Problem

"Up to now, research in the field has been slow due to competing theoretical and methodological paradigms…there is a need to go behind the general achievement patterns and open the black box of teaching and learning practices."

-- Klette, 2017, p. 1

"[The potential of classroom observation instruments] is dependent on observers [raters] who are carefully trained and supervised to provide accurate and consistent scores. [This results in] validity challenges stemming from rater error, which often remains high despite extensive efforts to train, certify and monitor raters."

-- White, 2018, p. 492



Background and History

Early systems: Variations of "4S" criteria: *Schooling, Seniority, Shoeshining, Sycophancy* (Olson, 1975)

1. Striking definitions/criteria to assess *teaching* (e.g., proper attire, shoe cleanliness)
2. Multiple evidence sources (e.g. observations, interviews, notebooks, lesson plans, seating charts)
3. No concern with technical properties (e.g., measurement error, reliability, validity)

A lot of progress made since:

1. Growing efforts to collect systematic evidence in classrooms
2. Attention to issues of construct definition
3. Quantification (*counting* vs. *rating*)
4. Qualifications of observers
5. Measurement error

For historical overviews. see Kasper et al., 2022; Martínez-Rizo, 2016.

“The concern for observing teacher behavior in the classroom (for control, research, or instructional improvement purposes) is again emerging as one way of attaining educational accountability . . . The people charged with the tasks of observing, assessing, and judging classroom teaching behaviors can avoid the pitfalls of previous observers if they take a close look at the historical development of observation procedures and reasons for utilizing teacher observation instruments.”

-- Lamb & Swick, 1979

1990s-2020s: The Latest Wave

- Increasing reach, aims, and intended policy uses

- Growing technical sophistication and research base
  - Conceptual frameworks and models of teaching
  - Instruments (e.g., CLASS, FFT, MQI, PLATO, ICALT, 3 Dimensions, TALIS)
  - Strong "content validity" of domains/dimensions
  - Psychometric methods/models (scoring, reliability, and validity)
  - High-profile empirical studies (e.g., MET – *Measures of Effective Teaching* [U.S.], *Pythagoras* [Germany], *TALIS Video* [International])
  - Validation efforts of observation measures within formal accountability systems (U.S.)

Classroom Observation: Multiple Purposes and Contexts

- **Theory Development**: Understand teaching practice; its nature, correlates and influences, effects and mechanisms
- **Comparative Research:** Measure and characterize instruction (OTL) across groups of students, localities, or countries
- **Accountability:** Support high-stakes judgments about quality/effectiveness of instruction (at the teacher or school level)
- **Evaluation/Institutional Development:** Monitor the implementation and effects of instructional programs; instruction as outcome and mediator
- **Professional Development:** Support lower-stakes judgments and feedback to teachers to improve practice

-- adapted from Correnti & Martínez, 2012

We Have Learned a Good Deal About Observing Teaching

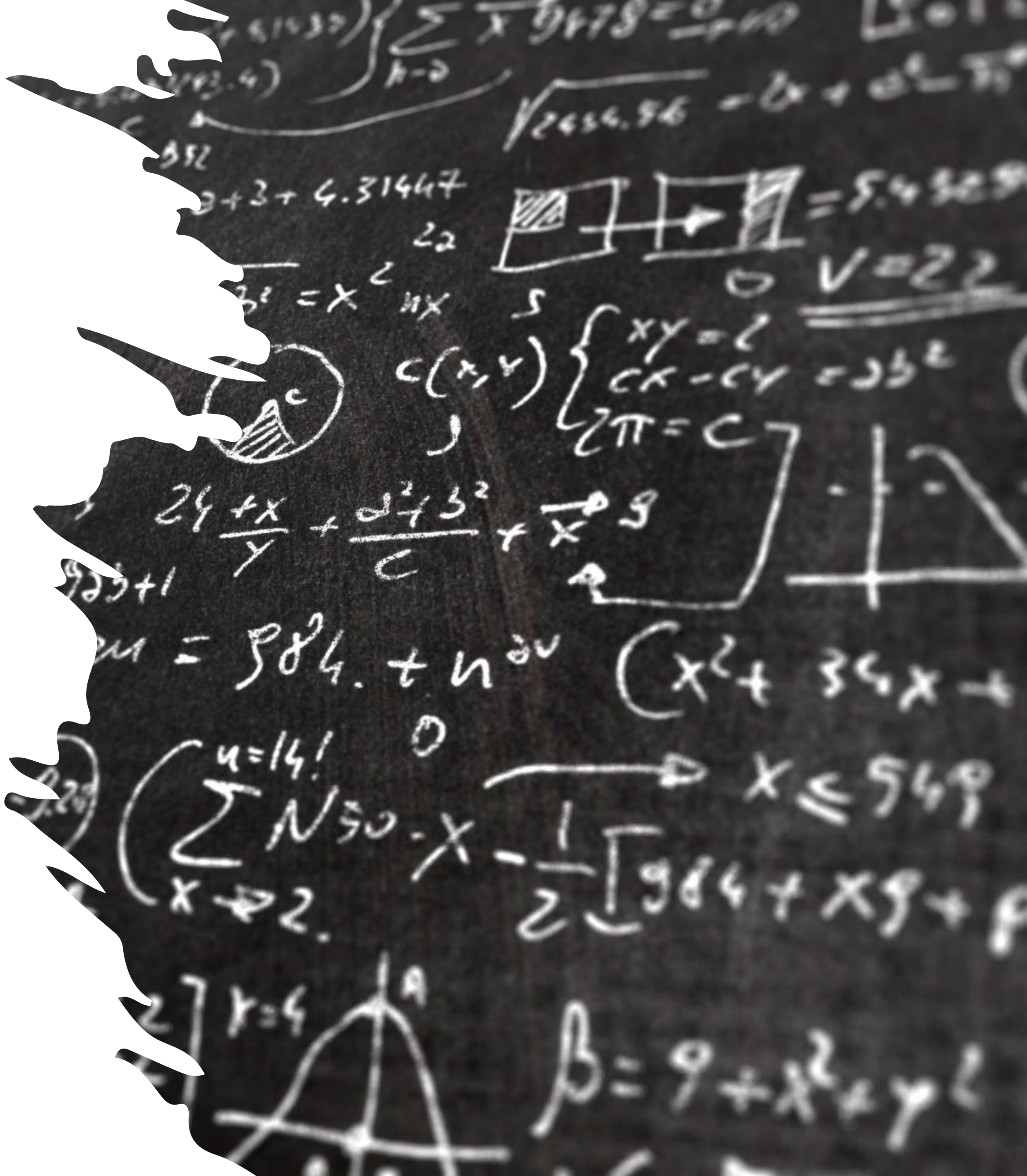
- Strengths and weaknesses of general and subject-specific protocols
- Characteristics of good raters (e.g., well-trained, experienced administrators vs. peers)
- Relative difficulty of scoring different dimensions (e.g., classroom management, teacher questioning, cognitive depth)
- Quality of scores under different observation methods (e.g., live observation, audio/video)
- ✓ Psychometric properties of different types of protocols and scales (e.g., checklists vs. subject-general ratings vs. subject-specific ratings)
- Key sources of error under different protocols, scoring designs

The Dominant Goal: Stable Estimates of Teaching Quality

State of the Art?

Sobering findings from recent well-designed studies:

- Low reliability - MET: 0.3-0.5 for one observation (0.5-0.7 for four); TALIS: 0.3-0.6
- Large standard errors (confidence intervals of 1-2+ points in 4-5-point scales)
- Large error variation across lessons within classrooms (5% to 40% of all variance)
- Low predictive and concurrent “validity” coefficients
- Little alignment to developmental trajectories of teaching
- Potential for significant inconsistency or bias related to student composition, teacher characteristics, grade levels, etc.
- High cost, logistical and practical challenges, limited utility informing feedback to teachers



Observations in Practice: Scores With Questionable Meaning

The results described previously for controlled studies can be safely expected to be a best-case scenario compared to use in real operational settings.

- Much wider range of schooling settings and contexts: Distribution of scores can be substantially different compared to research settings.
- Inconsistent to minimal quality control (e.g., training, qualifications, double scoring, number of observations, etc.)
- Observers are not disinterested parties and juggle various considerations in assigning scores.

Pursuit of Measuring Stable Traits of Teaching Quality Through Observation Remains Elusive

- Teaching is more complex than ever
- Methods and techniques are substantially refined but…
  - evidence is still inconsistent at best;
  - exhausted this particular technology; and
  - the path and likelihood of improvement is uncertain.

Our Thesis

- It was unrealistic to expect that these kinds of observation protocols would have produced more reliable and *valid* scores.
- The greatest promise of observational methods is to contribute to improving teaching through professional and institutional development.
- The *scientific* pursuit of understanding teaching through observation cannot be separated from the policy/political dimensions of its use.

Why Can't We Do Better?

The Challenge of Codification

"In methodological sections of classroom studies, and even in introductory publications about how to conduct research, the process of coding is sometimes described as a straightforward sorting of data into different categories. However, coding is more than a way of sorting data; it is a transformation of data from one form into another, wherein certain perspectives are bound to be systematically foregrounded and emphasized while others are not. Thus, codification changes not only how the data are organized into smaller entities but also how they are perceived as a whole."

-- Klette & Blikstad-Balas, 2018, p. 131

Considering Observation as a Problem of Categorization

- Humans make sense of much of the world by organizing things into categories (schemas, prototypes, scripts) (Focus of philosophers and psychologists)
- Categories share essential features, elements, structures etc.
- Natural categories (with physical features) are easiest to process, and most easy to get agreement on (e.g., colors, shapes, animal species)
  - o Even with natural categories, less good instances of a natural category are harder to process
    - Is a robin a *bird?* *vs.* Is a penguin a *bird?*
    - Is a dog a *mammal* vs. is a bat a *mammal*?
- Non-natural, social or human categories are much less obvious and difficult to achieve consensus (e.g., beauty, personality, poverty, type of government)
- We often organize the world of social experience through stereotyping - simultaneously adaptive and problematic

Perception Governed by Goals and Experience

-- Jerome Bruner

- *Organizing facts in terms of principles and ideas from which they may be inferred is the only known way of reducing the quick rate of human memory loss.*
- *Grasping the structure of a subject is understanding it in a way that permits many other things to be related to it meaningfully. To learn structure, in short, is to learn how things are related.*

The Inevitability of Inconsistent Categorization

- We ask observers to engage in multiple categorization decisions:
  - Identifying and recognizing evidence
  - Assigning evidence as relevant to the category(ies) of particular dimensions
  - Considering evidence as belonging to a particular score point category
- Protocols include:
  - categories at all levels invented by the protocol designers;
  - categories that are not natural in the philosophical sense; and
  - requirements that observers develop understandings of features as well as categories.
- But observers and teachers have already developed complex knowledge structures about teaching over many years and many experiences!
- Is it reasonable to expect any protocol and short-term training to lead to a fundamental restructuring of knowledge?

High- vs. Low- inference Indicators

Low-inference: Concrete, *objective* features/instances of behavior

- *Examples*: number of students called on, number of times teacher asked questions, number of minutes of independent work
- Unambiguous *categories* yield high reliability.
- Insufficient to capture quality:
  - But can map to/operationalize broader traits
  - Relates directly to student experience
  - Easier to influence and modify

High-inference: Abstract properties or qualities of practice

- *Examples*: "Cognitive challenge," "Investigating scientific questions," "Equitable participation"
- Also known as the key idealized traits of instructional quality that we ultimately aim to influence
- Quantification involves subjective judgement often resulting in lower reliability.
  - May be inconsistent with student perspectives and experiences
  - Low replicability for research/theory building
  - Harder to teach/influence in the short term

Moving Forward: So where is that vast sea of possibilities, then…?

- A highly sophisticated observation system providing reliable and valid data for high-stakes uses **is possible in theory**, but the investment required does not appear justifiable or sustainable compared to a formative system for producing desirable improvements in teaching.

- The greatest promise of systematic classroom observation is its potential to contribute to improving teaching through professional and institutional development by engaging in **long-term engagement with protocols** to:
  - promote common conceptualizations, models, and even language around teaching;
  - support individual and collective reflection and growth around teaching;
  - inform curriculum and instruction in teacher preparation programs; and
  - inform professional development and supervisory structures and programs seeking to offer pertinent, actionable, and contextualized feedback to teachers.

The Need for New Approaches

We offer two broad and potentially complementary approaches for moving our work forward.

- **Reconceiving measurement error**: Many facets of what has been conceptualized as measurement error can be rich sources of information critical to understanding and engaging in the support and improvement of teaching.
- **Using AI and advanced technologies**: Along with commercial hype, there are potentially productive uses of technology to support professional and institutional development.

Reconsidering Error: Learning Through an Interpretive Lens

1) Validity without reliability (Moss, 1994)

- Occasion variance is signal, not error: Instruction varies across lessons in a unit.
- Rather than *sampling* occasions, feedback for improvement calls for focusing on particular teaching activities – an interpretive lens can help understand how practice takes shape around the goals of a particular instructional event.
- Rater variance reflects differences in training and attention but also in perspective/expertise.
- Collaborative discussion helps with a) score agreement; and b) developing richer understanding of teaching for teachers and observers.

Reconsidering Error: Learning Through an Interpretive Lens (cont.)

2) Reliability without "reliability" (Mislevy, 2004)

- Reliability (like validity) is a property of inferences, not instruments or numbers. There can have *high reliability* for some inferences, low for others.

- Contrast reliability and value of average and consensus scores are *useful analogs in academia.*

3) Validity without "validity coefficients" (Martínez & Gitomer, 2024)

- Observation data with multiple dimensions and raters challenge standard psychometric models (low reliability, attenuated predictive coefficients, multidimensional x-classified structures).

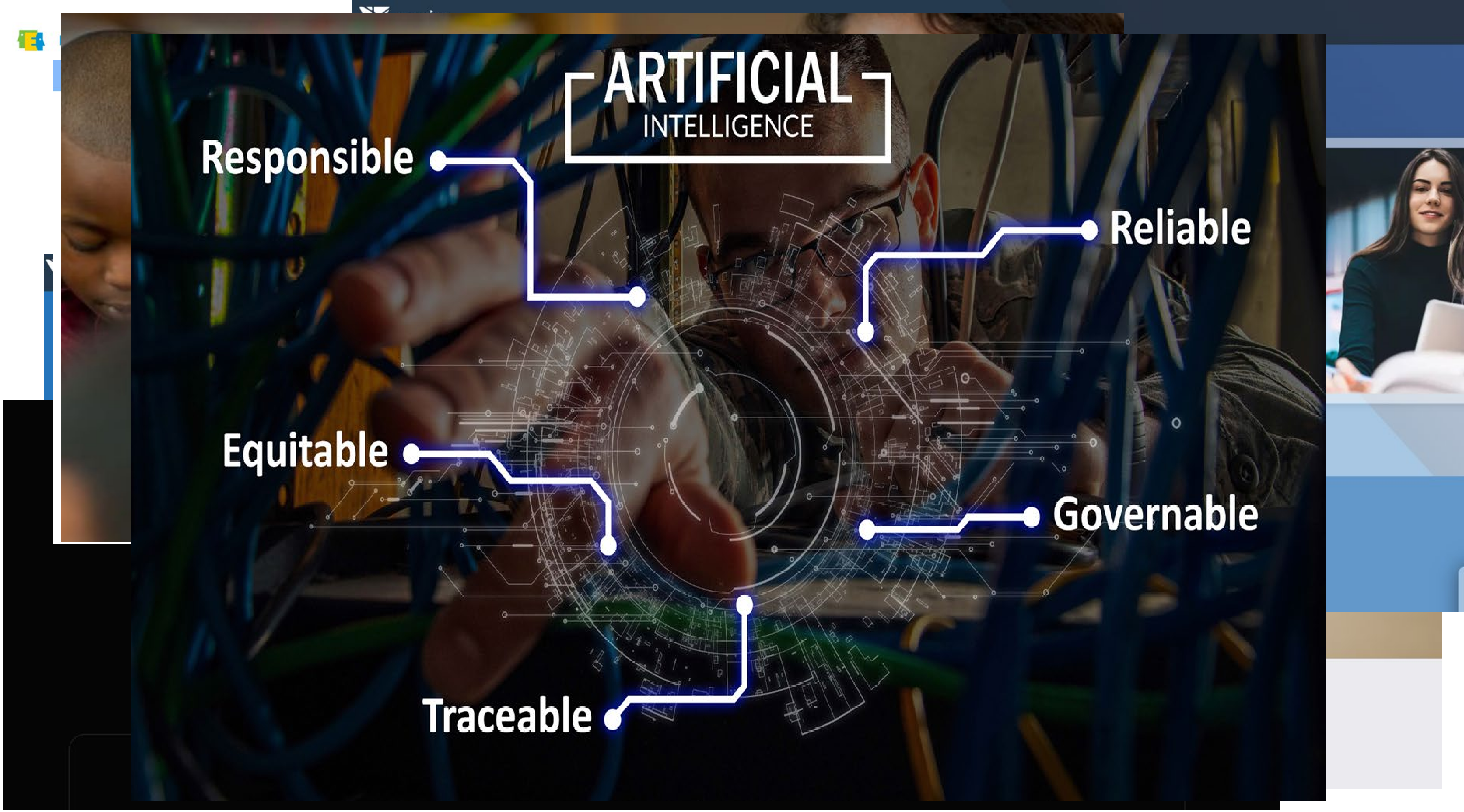
- As important, or more important, is content validity, face validity, use or adoption, and utility.

Artificial Intelligence (Re)visited

"One thing is clear to us, and that is that we won't be able to magically fix the kinds of complex conceptual, methodological, and practice challenges we are discussing here by sprinkling magical AI fairy dust…"

-- Gitomer & Martínez, 2023

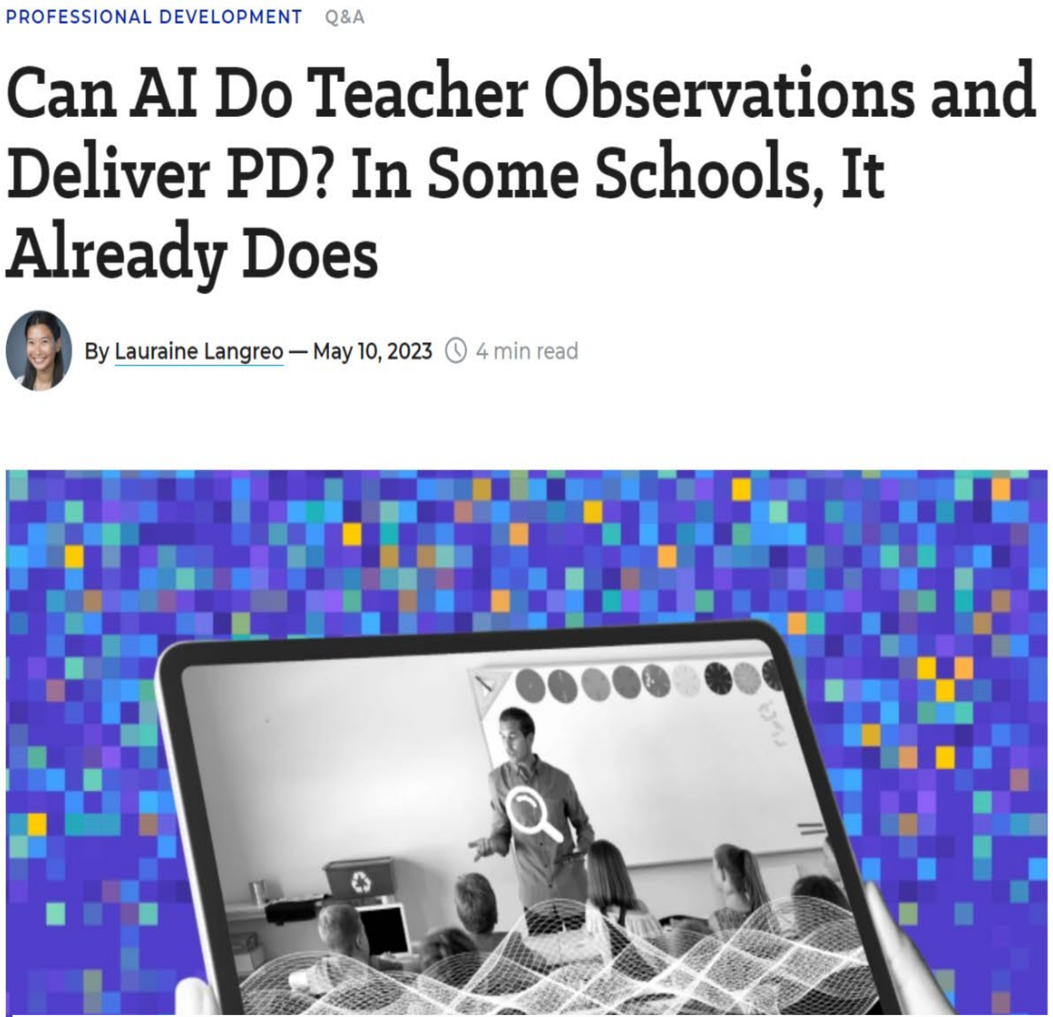
As it turns out, we were wrong…


ARTIFICIAL
INTELLIGENCE
Responsible
Reliable
Equitable
Governable
Traceable

Artificial Intelligence (Re)visited (cont.)

- What is “it”? A myriad of applications being marketed to districts in the United States.
  - Select video for teacher self-observation (or parent review)
  - Advanced, guided video tagging
  - Audio/video transcription and analysis
  - Private and adaptive coaching
  - …(a host of nebulous *others*)…
  - Automated “scoring” of classroom interactions
- A gold rush is under way. Results will be, in part, predictable…
  - Marketing running way ahead of the evidentiary basis
  - Exploration, concern, frustration, persistence
- But we also see significant promise. Some of these could even be useful, powerful, complementary tools for formative classroom observation.

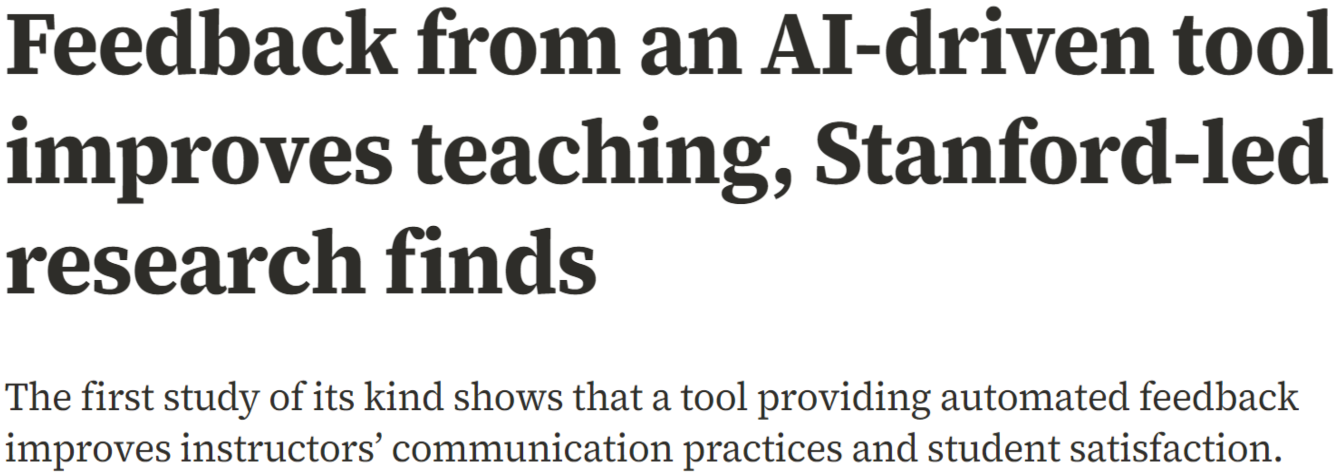
For example, **you can ask [tool]: “Which domains of Danielson's Framework For Teaching are used?”.** The AI will complete the analysis and give you an answer that includes timestamps that you can click on to review.



PROFESSIONAL DEVELOPMENT Q&A

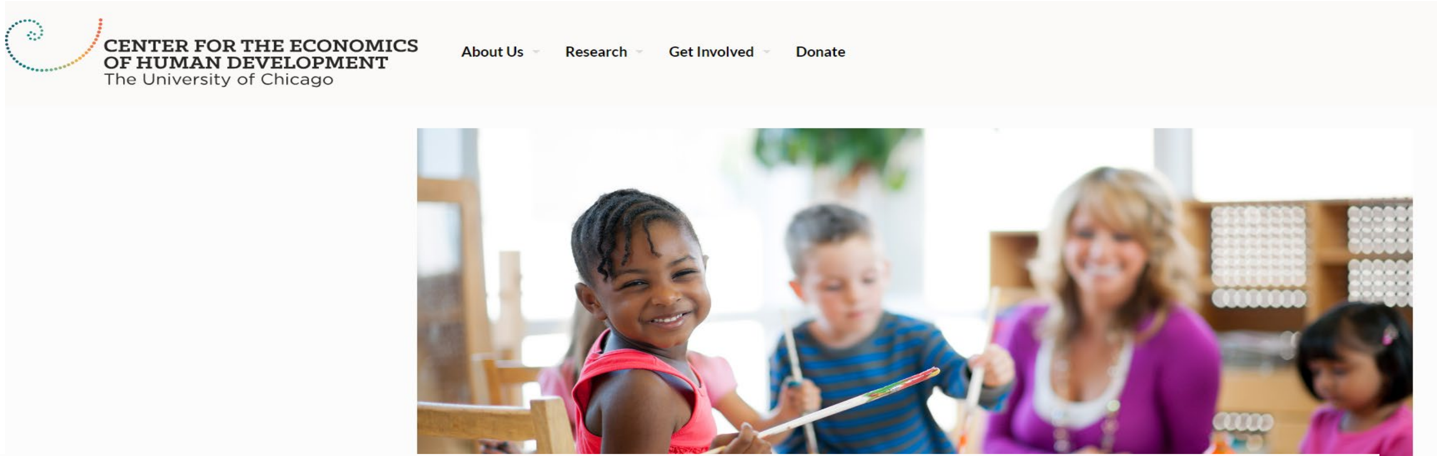
Can AI Do Teacher Observations and Deliver PD? In Some Schools, It Already Does

By Lauraine Langreo — May 10, 2023 4 min read

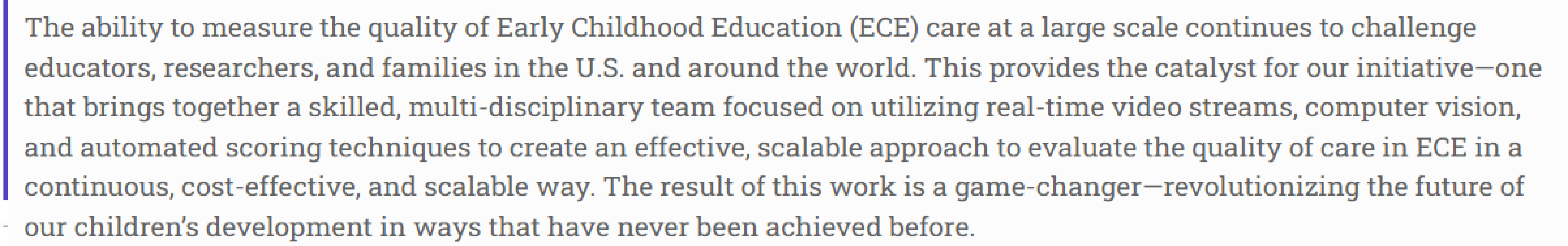


Feedback from an AI-driven tool improves teaching, Stanford-led research finds

The first study of its kind shows that a tool providing automated feedback improves instructors' communication practices and student satisfaction.



CENTER FOR THE ECONOMICS OF HUMAN DEVELOPMENT
The University of Chicago



The ability to measure the quality of Early Childhood Education (ECE) care at a large scale continues to challenge educators, researchers, and families in the U.S. and around the world. This provides the catalyst for our initiative—one that brings together a skilled, multi-disciplinary team focused on utilizing real-time video streams, computer vision, and automated scoring techniques to create an effective, scalable approach to evaluate the quality of care in ECE in a continuous, cost-effective, and scalable way. The result of this work is a game-changer—revolutionizing the future of our children's development in ways that have never been achieved before.

Artificial Intelligence (Re)visited: Areas of Promise

Automated measurement and monitoring of low-inference indicators and proxies of high-inference constructs

- Low-inference examples: classroom discourse, student engagement
- High-inference examples: cognitive challenge, classroom assessment

Supplementary tool for formative analysis/reflection

- Selecting videos
- Transcription
- Efficient structures to support coaching
- Collaboration structures for Professional Learning Communities

A range of others TBD...

Observation Systems as Tools for Inquiry

Future research and development on classroom observation:

- Realistic, policy-aligned frameworks for validity and usefulness
- Focus on instructional improvement
- Build around theories of teacher learning
- New measurement models (occasion, rater, and context *error*)
- Productive interpretive activities around observation
- Investigate/exploit AI capabilities for high-level repetitive tasks

Political/Policy Dimensions of This Work

- The policy context is more complex than ever.
- The obviously political: Partisan politics are a driving force.
  - United States - Race to the Top (2009): Led to consequential evaluation systems in many states
  - Mexico (2012-16): National teacher evaluation system implemented, then eliminated
  - Key actors: Federal, state, and local government agencies, teacher unions
- The less obviously political:
  - Views on teachers and the educational system
  - Achievement (primarily in mathematics and literacy) as valued outcome
  - Focus on distinguishing and ordering individuals vs. communal improvement
  - Key actors: Universities, Foundations (e.g., Gates, NCTQ), non-profits, vendors, the *public*

Thank you!
(We wish we were there with you!)

drew.gitomer@gse.rutgers.edu
jfmtz@ucla.edu