Using Student Work for Educator Accountability: The Road from Authenticity to Validity

Henry Braun
Lynch School of Education
Boston College

RILS'12

Boston, MA
September 13 2012

Outline

Background on educator accountability

Approaches to validation

Validity of test-based indicators derived from VAM

Validity issues with using end-of-course tests

Validity issues with using student learning objectives

Recommendations

Current Landscape

- Many states are revamping their educator accountability systems
  - Incorporating student outcomes
  - Aggregating different types of evidence
  - Introducing meaningful rewards/sanctions

- Most states are adopting
  - CCSS
  - Aligned assessment batteries (PARCC, SBAC)
  - Challenging performance standards (CCR)

- States are struggling to survive the multiple transitions and devising accountability systems for ALL educators – including teachers of "non-tested grades and subjects"

Validity

- An accountability system should be judged with respect to the criterion of **consequential validity**:
  - What are the intended outcomes (goals)?
  - To what degree have the goals been achieved?
  - What are the costs/benefits (short- and long-term)?
- The ideal is to devise an accountability system that is **systemically valid**:

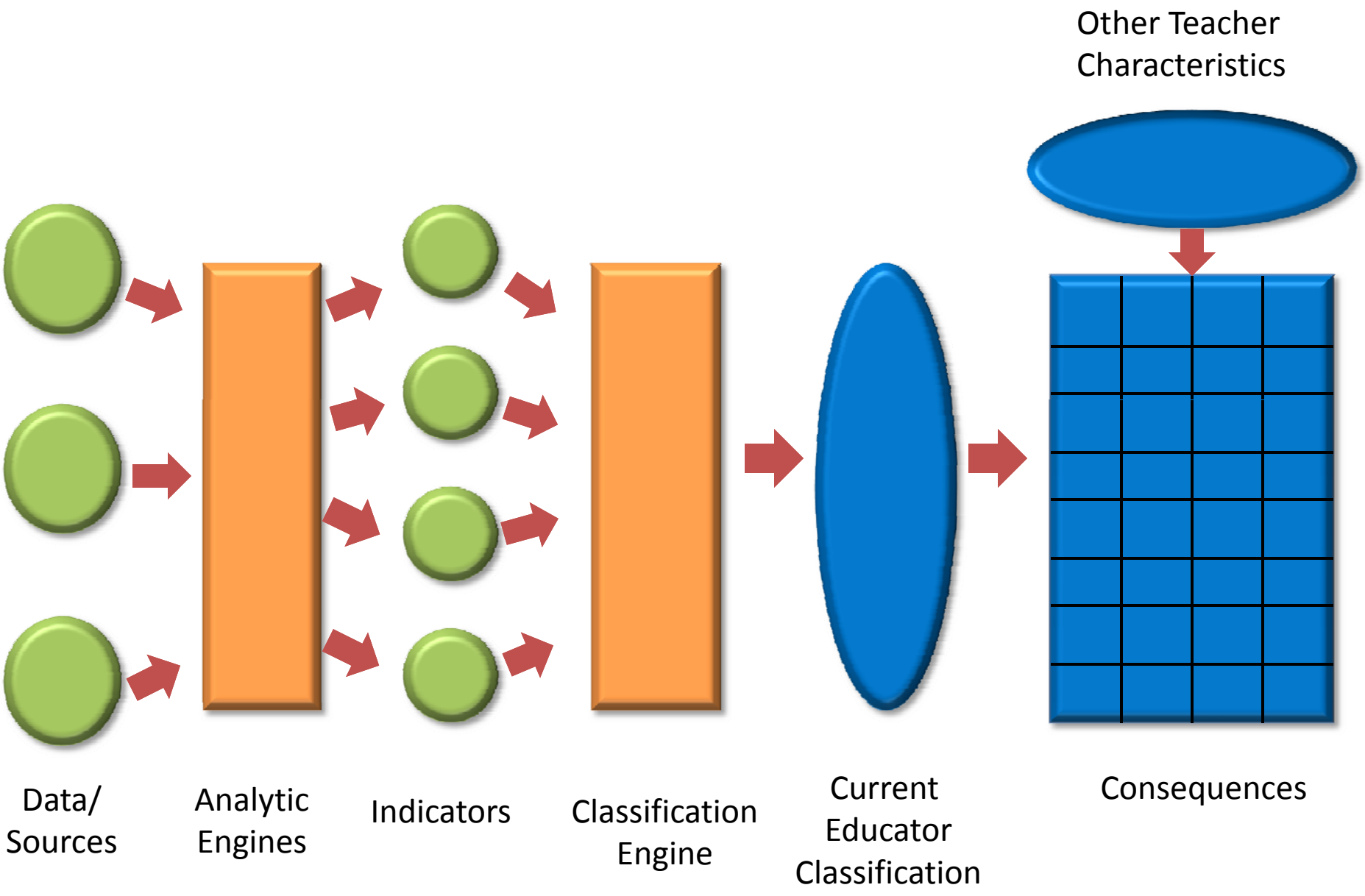
  It should contribute to the achievement of one or more of the intended goals – without undue deterioration in other aspects of the education process.

Goals for Educator Accountability

- Improve student outcomes through better instruction and related practices.

- Improve: Raise the bar, narrow the gaps
- Outcomes:
  - Academic/cognitive
  - Interpersonal
  - Intrapersonal

- Related Practices: Management of human capital that is
  - Effective and Efficient
  - Credible and fair (comparability)


Other Teacher Characteristics
Data/ Sources
Analytic Engines
Indicators
Classification Engine
Current Educator Classification
Consequences

Prospective Evaluation of Consequential Validity: Oxymoron or Challenge? (I)

Assessment ←→ System Design

Interpretive Argument ←→ Theory-of-Action

Construct Validity Argument ←→ Consequential Validity Argument

Prospective evaluation requires an elaborated **theory-of-action** that details the mechanisms by which the system will accomplish its goals – without undue "**collateral damage**".

The structure of the validity argument and the kinds of evidence needed in the two settings are quite different.

Prospective Evaluation of Consequential Validity: Oxymoron or Challenge? (II)

Methodology: Apply a logical combination of theoretical results, empirical findings and historical evidence to evaluate:

- The plausibility of the theory-of-action
- The ways in which the accountability system supports (or not) the theory-of-action

General Design Considerations:

- Components with appropriate operating characteristics
- Coherence and resonance
- Flexibility for adaptation and revision

Battle of the Indicators (I)

Test-based Indicators

vs.

Practice-based Indicators

How can we judge the contributions of these indicators to the consequential validity of the accountability system?

The argument must be wide-ranging – from relatively narrow technical concerns to broad educational issues.

Tracing the impact of specific components in a complex system is not a simple matter – and may not be feasible.

Test-based Indicators (VAMs)

PROs

Direct measure of an outcome

Face validity – objective measure

Evidence of causal linkages

Promotes educators' focus on valued outcomes

CHALLENGEs

Narrow band-width

Dependent on test quality & distributional characteristics

Lack of transparency

Reliability & Stability

Bias (non-comparability)
Lack of fit
Proxy predictors
Confounding factors

Little or no info for improvement

Need to set standards for decisions

•

Practice-based Indicators

PROs

Traditional

Direct link to teacher

Broader band-width

Info for improvement

CHALLENGEs

Traditional

Reliability & Stability

Bias

Causal link to outcomes

Administrative burden

Need to set standards for decisions

Battle of the Indicators (II)

One way to view this battle is to see it as a debate over which approach has greater **authenticity**… But authenticity is only the starting point for a credible validity argument.

Authenticity: The quality of being authoritative, valid, true, real or genuine. (Websters)

For our purposes, we must distinguish between **authentic** and **valid**.

Authenticity is a property of the measurement (indicator).

Validity is a property of the inferences made on the basis of the measurement (indicator) – and requires elucidation of the evidential contribution of the measurement (indicator) to the validity argument.

Authenticity

"The concepts of authenticity and directness (of performance assessment) are tantamount to promissory validity claims that they offset, respectively, the two major threats to construct validity, namely, construct underrepresentation and construct irrelevant variance.

(Messick, 1994)

… So the road from authenticity to validity may neither be short nor smooth!

Possible Potholes in the Road

- Bounds on authenticity can result in
  - Weaker relevance
  - Perverse incentives (goal distortion)
- Evidential contributions can be limited by
  - Weaker relevance
  - Plausible alternatives to initial warrants for causal linkage
  - Statistical properties
  - Imbalance between strength of evidence and proposed uses (consequences)

Validity of VAM estimates – Intro

Generating estimates of teachers' relative effectiveness is a type of measurement process and, hence, should be examined according to the guidelines set out in the AERA/APA/NCME standards.

(Hill, 2011; Braun, 2013)

This is a multi-phase process in which test scores of individual students are obtained, transformed, fed into an analytic engine and then aggregated to generate the desired estimates.

Validity strategies for such compound processes are not well worked out.

Validity of VAM estimates -- Claims

- Relevance
  - Based on measures of valued student learning
  - Causal link to teacher effectiveness
- Utility
  - Measurement properties at least as good (better than?) alternatives
  - Substantial variation among estimates and some fraction of teachers can be statistically distinguished from the “average teacher”
  - Can inform personnel decisions
- Consistency/Complementarity
  - Consistent with how education system performance is evaluated (by the public, others)
  - Appropriate balance to historic focus on practice-based indicators

Validity of VAM estimates – Threats I

- Test quality for the individual student
  - Construct underrepresentation
  - Construct irrelevant variance
- Test quality in the aggregate
  - Distributional characteristics
    - Floor/ceiling effects
    - SEM or classification errors
  - Relationships across grades
  - Transformations (preparatory)
    - Rescaling
    - Vertical linking

Validity of VAM estimates – Threats II

- Weak causal link due to multiple sources of bias
  - Poor model fit
  - Proxy predictors
  - Residual confounding
  - Class effects → Teacher effects (?)
  - Contextual complications
- Reliability/stability/sensitivity
  - Relatively high levels of volatility
  - Dependence on type of test and transformations used
  - Dependence on VAM used

Validity of Practice-based Indicators

- Based on measures of teacher practice and professionalism with a *prima facie* case for authenticity – but that can be challenged in a number of ways (depending on design and implementation of data gathering)
- Evidential contribution to consequential validity depends on
  - Causal link to student learning
  - Contribution to teacher capacity
- Research literature offers weak support for this link (exception: Hill et al. math study, 2012)
- MET study?

Implications for Accountability System Design

- Benefits of employing various types of indicators
  - Different targets
  - Fallible (bias and error)
  - Problematic causal linkages (of various kinds)
- Aggregation rules should respect both statistical properties and evidential complementarity
- Need for audit systems at multiple levels

End-of-Course Tests - I

Problem: How to transform student results into a TBI.

Challenge: Conventional VAM/SGP route not feasible because of multiple course sequences.

Solutions (conditional status models):

- VAM that accommodate missing predictors (by design)
- VAM that only use common assessments, (possibly) demographic information, class and school characteristics

End-of-Course Tests - II

Same PROs and CHALLENGEs as before, but ….

- Operating characteristics less well studied
- Tests of differential quality (district-level)
- More bias (?)
  - Poorer fit
  - Differential course taking patterns
  - Contemporaneous course taking

Student Learning Objectives (SLO)

- Teachers establish curriculum-based goals for
  - Individual students, or
  - Groups of students, or
  - Whole class
- Goals are based on some combination of
  - Prior achievement
  - Baseline measures
  - Other information`
- Student learning is measured in terms of success in achieving the objectives
- Summary statistic serves as indicator for accountability

SLO: A Shorter Path from A → V ?

- SLO can be viewed as a hybrid of test-based and practice-based indicators
- In principle, SLO can be developed for any context
- There generally should be a ….
  - plausible intuitive argument for authenticity
  - straightforward link from student performance to teacher effectiveness
  - reasonable expectation of positive impact on teachers' practices and professionalism

So what's the problem?

Interrogating Authenticity

- What does authenticity refer to: The SLO, the assessment of the SLO, both (or none of the above)?
- How should SLO be framed to support both instruction and assessment? What are some relevant discipline-specific considerations?
- To what extent do the SLO/A reflect the content standards?
- What are possible sources of construct-irrelevant variance?
- How should performance standards be established?
- Are there interpretive concerns with class-level SLO?
- What are the psychometric properties of the assessments?

Interrogating Evidential Contributions

- Comparability?
  - Within grade, across schools
  - Across grades, within schools
- Confounding factors?
  - Time on task
  - Extra-school inputs
- Impact on instruction?
- Impact on student learning (long-term)?
- Impact on teacher professionalism and satisfaction?

SLO Research Agenda

- Creating disciplinary frameworks for SLO to enhance authenticity and comparability
- Developing guidelines for high quality SLO/A
- Devising audit systems for SLO/A
- Statistical analyses
- Gathering evidence about impact on instruction and learning
- Gathering evidence about teacher advancement

General Considerations for Indicators (I)

Personnel decisions should be based on evidence that is

- Appropriate in scope
- Relevant
- Accurate
- Reliable

Sources of evidence should be examined critically (interrogated!) with respect to authenticity & evidential contributions to systemic validity.

Too often authenticity is treated as self-evident and evidential contributions are considered too narrowly

The validity argument should be an exercise in falsification and not confirmation (Popper, Cronbach, Messick)

General Considerations for Indicators (II)

- The practice of employing indicators based on student work for accountability should incorporate human judgment in a systematic manner for both QC/Audit and contextual interpretation
- Segmentation of indicator values into ordinal classes and aggregation of indicators needs greater methodological rigor
- The impact of the accountability system should be examined with respect to a broader range of student outcomes (systemic validity)
- The dynamics of human and system responses to high-stakes accountability in a time of transition should be studied through planned investigations and natural experiments

Conclusions

Test-based indicators should play a role in educator accountability – provided that they satisfy the demands of authenticity and evidential contributions to systemic validity.

Establishing that those demands are reasonably well met requires ongoing monitoring and research.

At the same time, we need to enhance accountability system design and implementation.