Viccissitudes of the Validators

Henry Braun

Lynch School of Education

Boston College

Presented at the

2008 Reidy Interactive Lecture Series

Portsmouth NH

Outline

- Accountability systems
- Validity and validation (general)
- Validation (specific)
  - Tests
  - Standards
  - Indicators
- Conclusions and prospects

Educational Accountability

- Hold the “system” (schools and teachers) responsible for the academic advancement of its students
- Externally mandated
- Monitoring of both inputs and outputs
- Variable consideration of context
- Differential incentives

Current Accountability Systems

- Goals
  - Raise learning for all students
  - Reduce achievement gaps
  - Improve system efficiency
- Emphasis on test-based outputs
- Minimal local context
- Standards-referenced indicators
- Simple and incomplete theory-of-action
- Meaningful consequences

Theory-of-Action

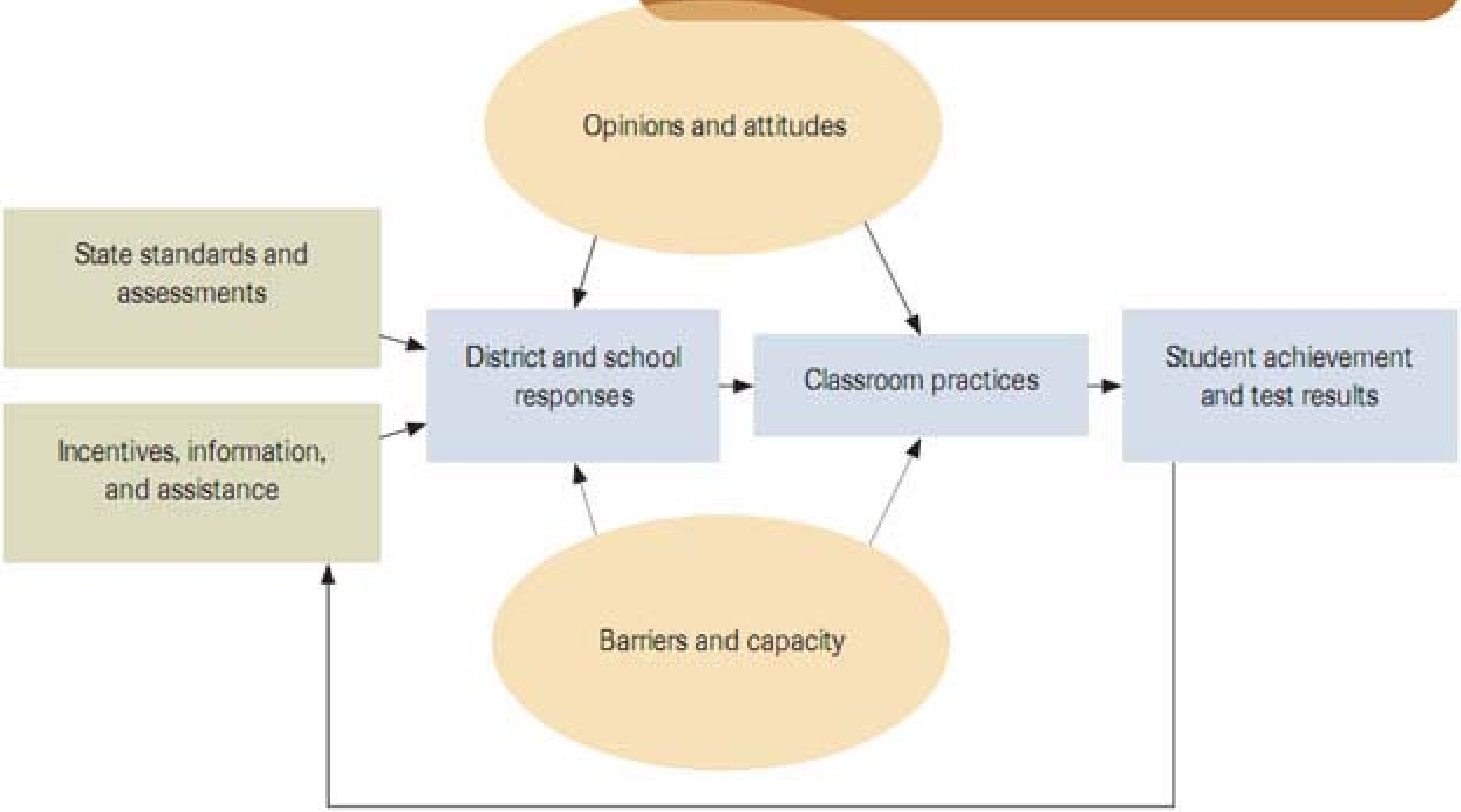
The justification for imposing a particular accountability system is the promise that it will accomplish the desired goals.

The mechanism by which this will happen is called the “**theory-of-action**”.

Too often it is stated in simplistic terms and ignores other (less desirable) behavioral responses that the system may elicit .

Standards-Based Accountability Theory of Action

The theory of standards-based accountability assumes an aligned system of district and classroom responses to instruct students to a state-prescribed set of skills. Student-achievement results then feed back into the system as a lever to prompt action at the district and local levels.



Opinions and attitudes

State standards and assessments

Incentives, information, and assistance

District and school responses

Classroom practices

Student achievement and test results

Barriers and capacity

Theory-of-Action (2)

Explication of a theory-of-action usually involves a high-level flow chart but rarely offers more detail on the processes that will operate.

It is precisely that level of detail that is required to develop a data collection plan to support a serious validation effort .... (More later)

Validity as a scientific enterprise

Validity is an ongoing argument that seeks to clarify what a measurement means and to understand the limitations of each score interpretation.

(adapted from Cronbach,1988)

- Validity as consideration of consequences

"Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores." (Messick, 1989)

Systemic Validity

Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (Access, Quality, Equity, Efficiency) within an education system, without causing undue deterioration with respect to other goals.

(adapted from Braun and Kanjee, 2006)

Validation

In validating an accountability system, the theory-of-action plays the same role as does the construct in test validation.

The more "ecological" view of validity embodied in the acceptance / embrace of consequential validity is perfectly suited to the questions raised about the impact of an accountability system on the education of our children.

Thus, consequential validity is the ultimate criterion by which we should judge an accountability system.

Validity Questions

Is the system working? If so, to what degree? If not, why not and what should be done about it?

The necessary evidence will comprise both quantitative and qualitative information.

The required data goes well beyond recording and analyzing students' test scores -- It requires tracking multiple facets of the system over time.

Such efforts are generally well beyond the capacity of state education departments and – in any case – are likely to be politically toxic.

STRONG STATES, WEAK SCHOOLS: THE DILEMMAS OF CENTRALIZED ACCOUNTABILITY (RAND, 2008)

- California, Georgia, Pennsylvania
- 70 superintendents, 260 principals, 2350 elementary & middle school math teachers
- Responses to high stakes accountability
- 95+ % variation among teachers (within schools)
- Conclusions
  - Variable linkage between central directives/expectations and teacher practices
  - Many principals and teachers lack requisite capacity to respond constructively to the new demands
  - Greater regulation will lead over time to greater uniformity
  - Not clear whether it will inspire innovation and creativity

Validation (2)

Study did not:

- Analyze links between teacher practices and student outcomes
- Investigate student trajectories
- Consider changes in overall resource allocation
- Examine patterns of teacher mobility

Ultimately, an evaluative judgment is a *tentative* causal conclusion based on *partial* evidence drawn from an *uncontrolled* study of schools and districts.

Caveats

"The chief fault of the testing movement has consisted in its emphasis upon content in highly academic material ... the fact that a particular pupil shows a marked improvement in reading or spelling may give some indication that a teacher is improving her performance ... but the use to which the pupil puts that knowledge is the only significant point in determining the significance of subject tests in measuring the educational system."

Ridley and Simon (1938, as quoted in Rothstein, 2008).

Caveats (cont.)

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

(D. T. Campbell, 1979)

"Distortion and Risk in Optimal Performance Contracts"

(George Baker, 2002)

Barriers to Validation

- Constraints
  - Fixed features (externally imposed)
  - Time
  - Cost
- Capacity
- Complexity
- Inertia
- Changing players

A Bottom-up Strategy?

- Begin by examining particular system components
- Consider degree of coherence among components
- Evaluate consequences (broadly conceived)
- Suggestions for redesign (in the small and in the large)

Standards for Educational Accountability Systems (CRESST/CPRE)

- System components
- Tests
- Stakes
- Public reporting formats
- Evaluation

Validating the Test

The lack of a gold standard means that the validation process must have an explicit strategy and rationale.

Kane (2004) suggests a two-phase approach:

Interpretive argument: Build a chain of reasoning from the test construction process to the desired claims.

Validity argument: Gather theoretical and empirical support for the “truthfulness” of the claims and to establish appropriate boundaries.

Validating the Test (2)

- Types of tests
  - NRT vs. CRT
  - End-of-course vs. Cross-cutting skills
- Threats to evidential validity
  - Construct underrepresentation
  - Construct-irrelevant variance
- Evidence regarding consequential validity
  - Student learning
  - Instructional practices
  - Administrative practices

Validating the Test (3)

The “test” is a “system” and should be evaluated as a system

- Design
- Development
- Materials preparation
- Administration
- Data processing
- Scoring
- Scaling
- Reporting

Validating Performance Standards

- Increasingly, academic performance is being communicated in terms of standards (e.g. 30% of students at or above proficient)
- Indicators framed in terms of standards are being used to track trends in differences and/or changes in academic performance
- Consequential decisions about students and/or schools are being made on the basis of results framed in terms of standards
- Policy-makers and the public make inferences about public schools based on their interpretations of the standards and standards-based reports

Validating Performance Standards (2)

“Arguments and procedures supporting a performance standard ... may differ according to the breadth of the claim the performance standards sets forth.”

“In practice, though, the performance standard always embodies a ... claim, pertaining to capabilities for performance in nontest settings.”

(Haertel and Lorie,2004)

Validating Performance Standards (3)

“Standard setting still can not be reduced to a problem of statistical estimation. Fundamentally, standard setting involves the development of a policy about what is to be required for each level of performance. This policy is stated in the performance standards and implemented through the cut scores.” (Kane, 2001, p. 85, emphasis added)

Performance standards are (usually) operationalized by means of a “cut-score” on the reporting scale.

Validating Performance Standards (4)

- Historically, standard setting has been a retrospective judgmental process carried out
  - independently of the assessment design process
  - after the assessment is administered the first time.
- The consequences of a retrospective approach are
  - Reliance on subject matter expertise rather than research on student learning and development
  - Tendency to conflate policy and psychometrics
  - Difficulty in achieving coherence of cut scores across grades
- Risks
  - Cut scores may not be well supported psychometrically
  - Insufficient evidence to adequately support desired inferences

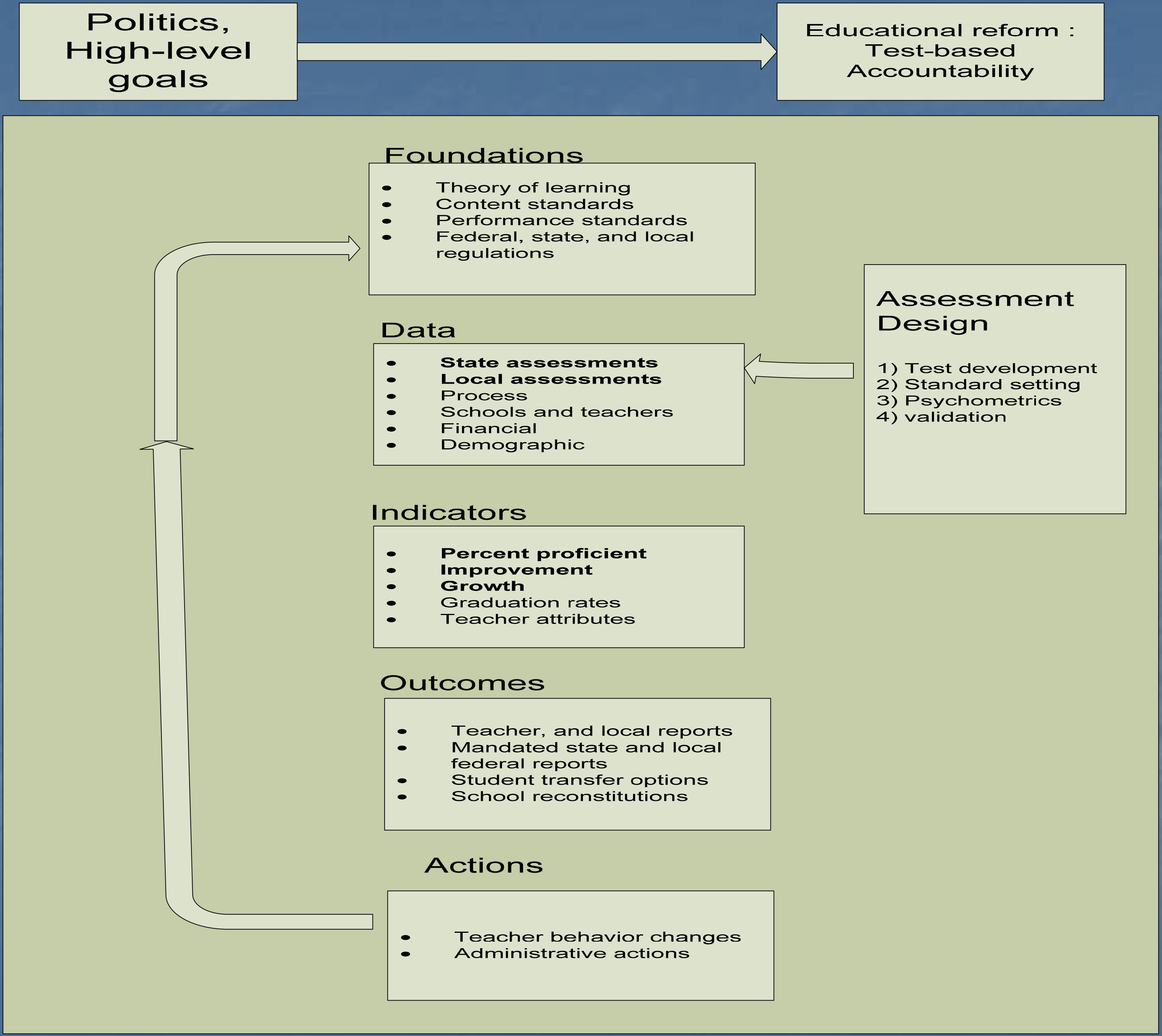
Validating Performance Standards (5)

- With no gold standard, “procedural validity” has been the touchstone for evaluating standards
- But the validity argument demands that we examine different aspects of “standards-in-use:
  - Credibility of interpretation
  - Classification reliability
  - Predictive efficacy (in-school, out-of-school)
  - Statistical properties of indicators derived from standards

There are important implications for establishing valid performance standards.

Validating Indicators

- Indicators are statistics calculated from student-level data that are used directly for decision-making
    - Percent of students exceeding the proficiency cut-score
    - The change in the percentage of students exceeding the proficiency cut-score
    - The percentage of students gaining at least 20 scale score points in math from 10th to 11th grade
    - Percent of 9th grade cohort graduating high school in four years
- There is an implicit assumption that the indicator validly captures an important aspect of system functioning – and that it is appropriate to evaluate the system (at least in part) on the value of that indicator



Politics, High-level goals

Educational reform : Test-based Accountability

Foundations

- Theory of learning
- Content standards
- Performance standards
- Federal, state, and local regulations

Data

- **State assessments**
- **Local assessments**
- Process
- Schools and teachers
- Financial
- Demographic

Assessment Design

1) Test development
2) Standard setting
3) Psychometrics
4) validation

Indicators

- **Percent proficient**
- **Improvement**
- **Growth**
- Graduation rates
- Teacher attributes

Outcomes

- Teacher, and local reports
- Mandated state and local federal reports
- Student transfer options
- School reconstitutions

Actions

- Teacher behavior changes
- Administrative actions

Validating Indicators (2)

The Janus Strategy

- Backwards:
  - Are the data underlying the indicator(s) valid for the purpose?
  - Is the construction of the indicator(s) defensible?
- Forwards:
  - Are the statistical properties of the indicator(s) consistent with the desired inferences?
  - Is the indicator set a reasonable basis for the intended decisions?

Validating Indicators (3)

General guidelines

- Indicators based on one source of data are very susceptible to corruption
- Technical analysis matters
- Lack of local context is problematic
- Strong incentives/sanctions not warranted

Validating Indicators (4)

Exemplars

- Percent proficient
- Trend in percent proficient
- Trend in difference in percent proficient
- Growth to a standard
- Absolute growth
- Estimate of value-added

Lessons Learned

- Accountability is expected and appropriate for any publicly funded enterprise
- Current accountability systems
  - **Are under-designed and over-hyped**
  - **Often substitute ideology for technical analysis**
  - **Ignore the dynamics of human responses**
  - **Too crude for the intended job**
- Experience in other fields (e.g. health care, law enforcement, business) tells us that it is rare to find an accountability system that works “as intended”

Lessons Learned (2)

"Whenever you try to legislate professional behavior, there are bound to be unintended consequences....Nor is it clear that pay for performance (P4P) will actually result in better care.

Doctors have seldom been rewarded for excellence, at least not in any tangible way... At first glance, P4P would seem to remedy this problem. But first its deep flaws must be addressed before patient care is compromised in unexpected ways."

[S. Jauhar, M.D., NY Times, 9/9/08]

Prospects

- Validation focused on components can make an important contribution
- In the long-term, design considerations must become more salient in the construction of accountability systems
- Because politics (almost) always trumps psychometrics, measurement personnel should try to engage more directly in the policy-making process

An accountability system that accomplishes flexible regulation in the service of constructive improvement of education outcomes is a rare beast indeed!

Coda

Validators → Gladiators