Improving the Validity of Accountability Systems

Eva L. Baker

UCLA Graduate School of Education & Information Studies
Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing

RILS
October 2000

CRESST Partners

UCLA

University of Colorado

Stanford University

The RAND Corporation

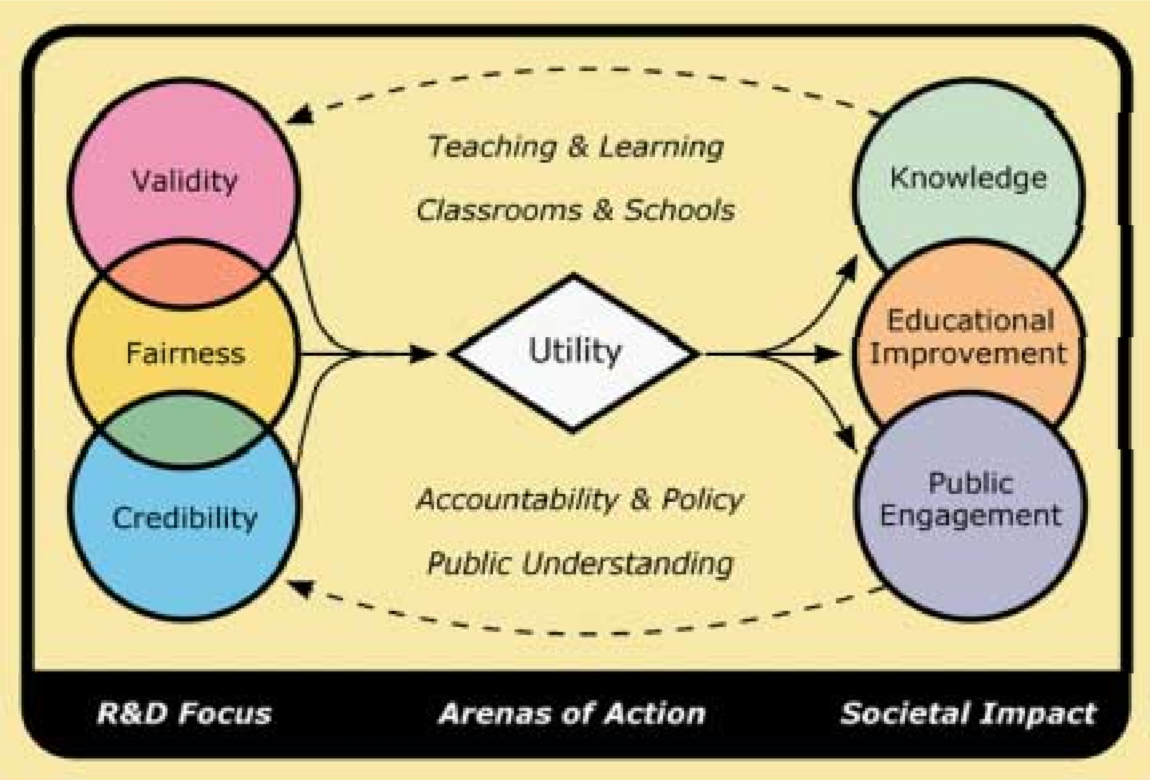
The University of Pittsburgh (LRDC)

Educational Testing Service

University of Cambridge

University of Southern California

CRESST Conceptual Framework



Improving the Validity of Accountability Systems

Why?

Accurate inferences essential

Guide system with increased credibility and tranquility

Assumptions

Validity strategy is independent of reform style

Requires a source of expertise to legitimate decisions

Comparisons among systems play a role

We know enough to do it

Present Approaches (1980-2000)

Influence legislation and regulations

Design guidance/self-interest

Ad hoc studies and analyses

Testimony in hearings and cases

Publication in the technical/public literature

Yield

Not great

Undermines credibility of the experts

Inappropriately politicizes the problem

Modest improvements

Options for the Research Community

More of the same

Models of accountability systems or stand-alone components

“Ideal” Systems, “Thought Experiments”

Public articulation and endorsement of standards for accountability systems

Develop evaluation models and guidelines

Accountability Standards

Goals

Audience

Design strategy

Early examples

Endorsement plan

Evaluation models

Goals for Accountability Standards

Clear guidance for design or analysis

Relevant to key audiences

Credible

Broad and stable template to evaluate system development

Goals of Accountability Systems

Improve student learning

Improve the quality of schools and districts

Improve overall system quality

Improve access to education

Monitor costs

Recognize responsibility

Motivate participants

Build public confidence in education

More Goals

Reduce the disparity of performance between identifiable subgroups

Improve economic competitive status among key states

Assess the effectiveness of key programs

Determine needs

Provide assistance

Provide transition to post-secondary options

Functional Goals

To raise test scores

To move indicator(s)

To show early progress somewhere

Arguments About Short-Term Focus

Shifts in attention

Shifts in leadership

Teacher quality shortfall

Motivate the public

How Is It Going?

Little evidence that long-term goals are being achieved

Controversy about test-prep, incentives

Questions about short-term gains, long-term effects

Turbulence

Just-in-time policy retreats

Standards for Educational Accountability Systems

Achievable set for guidance, reflection, and review

Public audience—legislators, media and educators

Based on research or best practices

Endorsements rather than consensus

Why This Audience?

Media controls public understanding of education

Media searches for serious roles

Media can mobilize the public

Media represents (along with donors) major pressure point for politicians

Principles Underlying Accountability Standards

A developmental perspective for the system

System information accurately represents the state of affairs

Indicators substantially under the control of the accountable institution or personnel, usable knowledge

Fair to all parties

Partial Precedent: *Standards for Educational and Psychological Testing*

(AERA, APA, & NCME, 1999)

Consensus of the field

Referenced in statutes, regulations, and case law

Guidance to developers

Requirements for users, test takers, and those with authority to mandate

Contrasts with Test Standards

Apply to systems of tests and indicators

Inferences apply to multiple levels, people/institutions

Knowledge base broader

Different structure and organization

Different time line

Proposed Targets of Standards: Emphases, Not Categories

System validity

Measurement

Accuracy and technical quality

Implementation

Special needs populations

Incentives, sanctions, stakes

Evaluating effects

Which Priorities?

Sources are legal and regulatory

Best practices

Empirical studies

Analytic, e.g., policy in conflict

Examples of Proposed Standards—Highly Derivative

Validity information is given for planned purpose(s) intended for a test or indicator (Validity)

Tests should minimize factors irrelevant to the domain assessed, e.g., language complexity with a science knowledge test (Special Populations)

High-stakes decisions should not be made using results of only one measure (Accuracy)

Examples of Proposed Standards (Cont.)

A test (or system) used to judge the growth of individuals, programs, or institutions should show evidence that change is the result of educational interventions, e.g., instruction (Validity)

When high-stakes decisions use multiple measures, the method of weighting all measures must be public (Incentives)

Conglomerate Example

Accountability systems should not be used for high-stakes decision making in the absence of adequate safeguards assuring accuracy and fairness. These safeguards should include multiple opportunities for test taking, adequate notice, evidence of instructional sensitivity, adequate OTL, and multiple measures.

Prospective or Retrospective Emphasis?

General

When do we rely on prior evidence, e.g., supported validity claims, and when do we attempt to collect evaluation data in support of validity claims?

Does evaluation focus mostly on side effects?

What is the basis for selecting evaluation questions?

Evaluation Variations

The validity of test results used for accountability decisions should be subject to ongoing evaluation.

The validity of system inferences about student learning and program effectiveness should be subject to ongoing evaluation.

The validity and consequences of accountability systems should be systematically and periodically evaluated.

Studies evaluating effects should be conducted to determine the degree to which the system:

Supports high-quality instruction

Promotes student access to education

Minimizes corruption

Affects teacher quality

Produces unanticipated outcomes

Cost

Quality at what cost?

Compared to whom?

Over what period?

When schools are placed in categories based on student assessment results, the likelihood that they would be placed in a different category if another class of students had been tested on a comparable version of the test needs to be provided.

Comment:

There is a widespread belief that tests yield more precise information about students and schools than they actually do. Hence, it is important for accountability systems to provide users with information about the precision of the results that are reported.

The standard is a natural counterpart in the context of school accountability systems to the position articulated in the *Test Standards* for student use of test scores to classify students. According to the *Test Standards*, where student results are reported in terms of a small number of performance levels such as Advanced, Proficient, Partially Proficient, and Unsatisfactory, the results need to be accompanied with reports of how likely a student who is placed in one category actually belongs in another, or how likely it is that a student would be placed in a different category if that student were retested with a comparable version of the test.

When results for students are reported in terms of scale scores or percentile ranks, indicators of the precision of scores should be provided. The *Test Standards* could be satisfied by reporting either the standard error of measurement or the likelihood that a student with a given score would receive a score that differed by a specified amount if a comparable form of the test was administered. When school results are reported in terms of average scale scores or the percentile rank of the average, the comparable school accountability standard would require that an indication of the error be provided, using the standard error or the likelihood that the average would differ by a set amount if a comparable form of the test was administered to another class of students in the school. (Accuracy)

Accountability systems that lead to decisions about or categorizations of schools or progress need to use multiple approaches in assessing achievement in order to increase the validity of the decisions or categorizations.

Comment:

Both the *Test Standards* and the ESEA Title I legislation call for the use of multiple measures. They do so in different contexts, however, and with somewhat different purposes in mind. The focus in the *Test Standards* is on use of tests to make high-stakes decisions about individuals.

In that context, the call for multiple measures first refers to providing students several opportunities to retake tests that might be used for decisions such as the award of a high school diploma. This requirement is intended to reduce the likelihood that students who should pass actually fail due to the fallible nature of tests. It also allows students to improve their knowledge and skills before retaking the test. The second use of multiple measures in the *Test Standards* refers to the notion that students should be allowed to demonstrate their achievement in more than one way, e.g., through essay assessments or projects as well as a standardized test. This second meaning of multiple measures comes about because it is recognized that different types of assessment can add to the validity of the decision. The use of multiple measures in the Title I legislation is more in keeping with the second use of the term in the *Test Standards*. Multiple approaches to assessment are encouraged as a means of enhancing the validity of the information obtained about student achievement.

Yet another use of the term multiple measures refers to the use of assessment to span a range of subject-matter areas. This latter interpretation is intended to encourage attention to instruction across the range of the curriculum. The multiple content-area emphasis applies to the use of results as summaries of school achievement as well as to individual students.

In keeping with the *Test Standards*, accountability systems that include high-stakes decisions about individual students should provide multiple opportunities to take the assessment in order to improve the overall accuracy of decisions. (System Validity)

Desirable Features of Accountability Standards

Appropriate for audience(s)

Common language but enough guidance

Parsimonious and high priority

Two-tier: Common language/technical?

Companion to Standards: Evaluation Models

Explicit designs to be broadly shared

Model questions and data collection

- *To focus attention on key issues*
- *To address scale*
- *To encourage comparative work*

Evaluation Models: External Use

To support state and local development of RFPs

To broaden the range of questions

To illustrate appropriate levels of effort

To suggest types of extant and special data to be collected

Evaluation Models: Grass Roots Use

Statewide but locally-sponsored efforts to evaluate accountability systems

Common questions, vetted data collection procedures

Coordinated reporting

External integration of findings for credibility

Needed from the R&D Community

Better conceptions of alignment

Better measures of instructional practice

Stronger guidance on validity of change

Evidence about incentives

Models of success with low-performing kids and schools

Needed from the R&D Community (Cont.)

Cheerleading and support for grass roots responsibility

Integration of findings from multiple sources within a state and for comparison purposes

Process

Generation

Interim release

Augmentation—revision

Partner and constituent input—state agencies and legislature, CPRE, ECS, professional organizations

Revision process

Endorsements

Immediate Next Steps

Suggestions for *Standards*

Opportunity to talk

Web contact

Continuous vetting

CRESST Web Site

http://www.cse.ucla.edu

Suggestions for Targets or Standards?