



*Driven by Policy: Validity and
Beyond*

R. Gene Hoffman, Ph.D.
ghoffman@humrro.org

CCSSO
National Conference on Student Assessment
June 15-18, 2008

An obligatory slide on reliability..

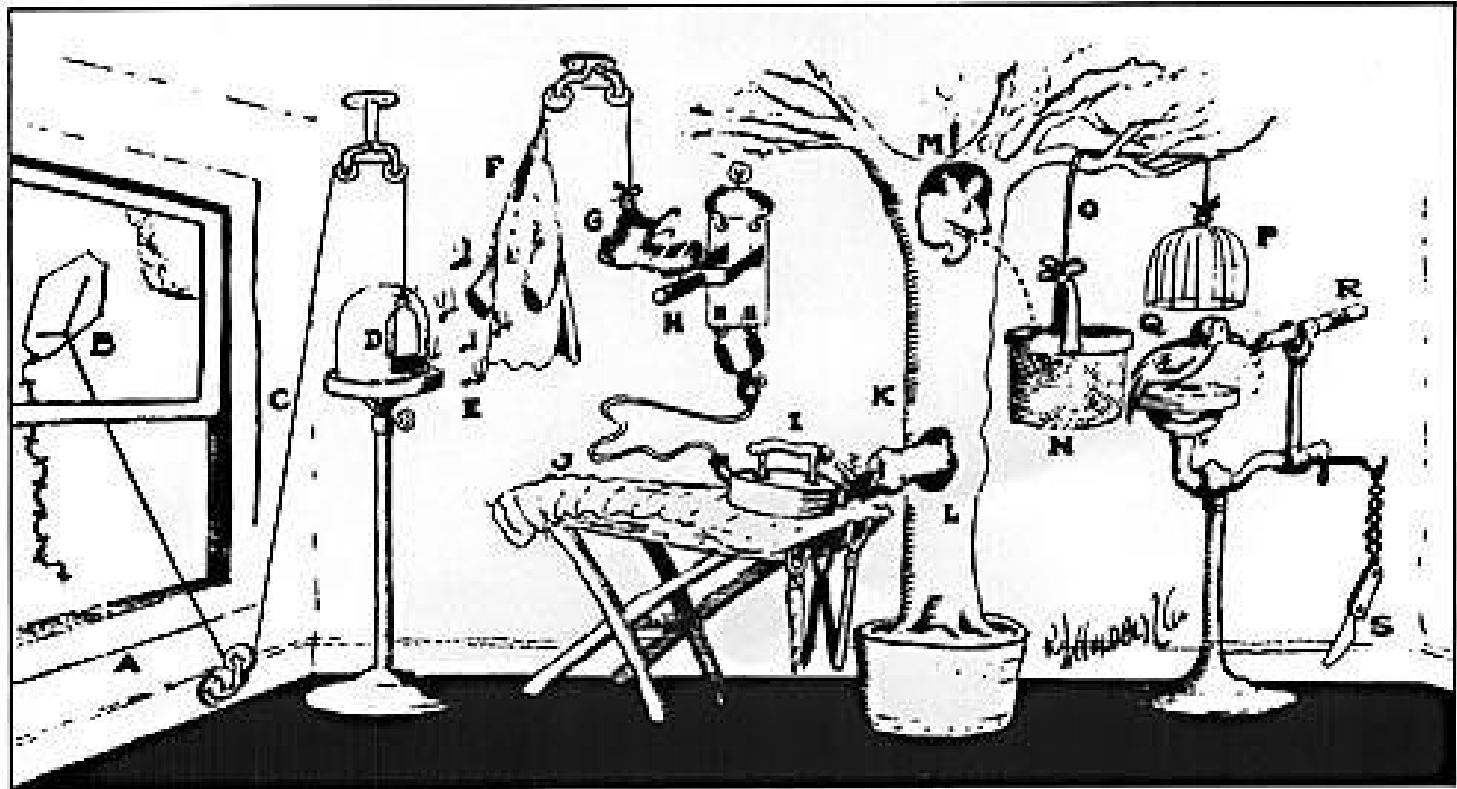
“When Bellows (1941) attempted to list criteria for a criterion [aka, tests], he included such criteria as reliability, accessibility and cost, and acceptability to the sponsor

...

These criteria possess a crude practicality... [that] seem to constitute an attempt to say. ‘We don’t know what we are doing, but we are doing it very carefully, and we hope you are pleased with our unintelligent diligence.’ “

(Wherry, 1957)

Nothing is ever simple...



Pencil Sharpener RUBE GOLDBERG (tm) RGI 038

Policy is bigger than validity

- ◆ Testing used as a lever for educational change
- ◆ Obviously, there is a need to address the impact that testing has on educational effectiveness.
- ◆ Other issues beyond validity:
 - Utility
 - Cost Tradeoffs
 - Local Value
 - System credibility
 - Publicly digestible communications

Validity's role

- ◆ *Validity* is a judgment about the usefulness of test scores for some particular, explicitly defined set of decisions or inferences.
 - A test with multiple purposes has multiple validity considerations
 - Inferences about students, teachers, schools, districts, states, nations
 - Normative issue is terms of scale scores and criterion references issues in terms of performance levels
- ◆ Accepting or rejecting the purpose of a particular use of a test is not a validity judgment.
 - Accepting or rejecting test purpose is a separate judgment that precedes judgments about validity.
 - A strategy of critics who do not favor high-stakes accountability as a reform lever is to use attacks on test validity to cloud debate on the acceptability of the purpose.

Utility

- ◆ Utility is bigger than consequential impact, because there is more to an accountability system than the test per se.
- ◆ From other perspectives, *utility* refers to the system-wide benefits of using the assessment and accountability system to meet program goals.
- ◆ Utility is not a simple function of test validity.
 - Assessments could be flawed (no test is perfect), but their implementation may stimulate positive changes in educational practices.
 - In spite of the testing flaws, an accountability program could have high utility – and the converse!
 - However, if the assessments are substantially flawed, we could expect the system to eventually fail.

Cost Trade-offs

- ◆ Performance appraisal systems naturally create stress.
- ◆ The occurrence of “negative” consequences must be placed in the context of utility.
- ◆ Trade-off must happen – within and between individual stakeholders
 - Individual Level: all actors
 - Organization: Schools, Unions, Universities
 - Constituents: Parents, Real Estate Agents
 - Political
- ◆ You can't please everyone all of the time....
 - Good pain versus bad pain.....

Local Value

- ◆ In addition to more global judgments about system utility and the trade-off of costs and benefits, *local value* is also an important criterion for evaluating the assessment and accountability system.
- ◆ If the timing and level of detail of test scores do not assist teachers, schools, and/or districts to make changes, then the system will be perceived as ill-conceived or irrelevant to local needs.
- ◆ Professional development and numerous student and teacher support systems come into play.

System Credibility

- ◆ All of these issues—validity, utility, costs trade-off, and local value combined with other issues, events, and hidden agendas—form a constellation of judgments that we can label *system credibility*.
- ◆ For a system to be fundamentally credible, system goals and objectives must be clear and accepted and the role of the test in meeting those goals and objectives must be clear and accepted.
- ◆ In addition, the system must be operationally sound; that is, as free of production and processing errors as conceivable. Such errors directly undermine credibility and can dilute validity, utility, and local value.

Publicly Digestible Communications

- ◆ Given that validity itself is complex, the additional judgments of utility, local value, and credibility make for an even more complicated picture. This situation underscores the need to highlight one more issue.
- ◆ *Publicly digestible communications* about evidence and judgments concerning all of these criteria must be readily available to all stakeholders.
- ◆ Misinformation and lack of information can be as detrimental to a program as low test validity. Therefore, accumulated evidence relevant to validity, utility, local value, and credibility must be indexed and accessible.
- ◆ Corollary – Expectations are tough to manage!

Sample Policy Goals

- ◆ Elevate teaching and learning to high standards
- ◆ Hold schools and districts accountable for improving learning.
- ◆ Ensure that learning includes core academic content, basic skills, higher order skills and application of knowledge.
- ◆ Focus attention on all students.
- ◆ Assist schools in improving their instructional effectiveness.

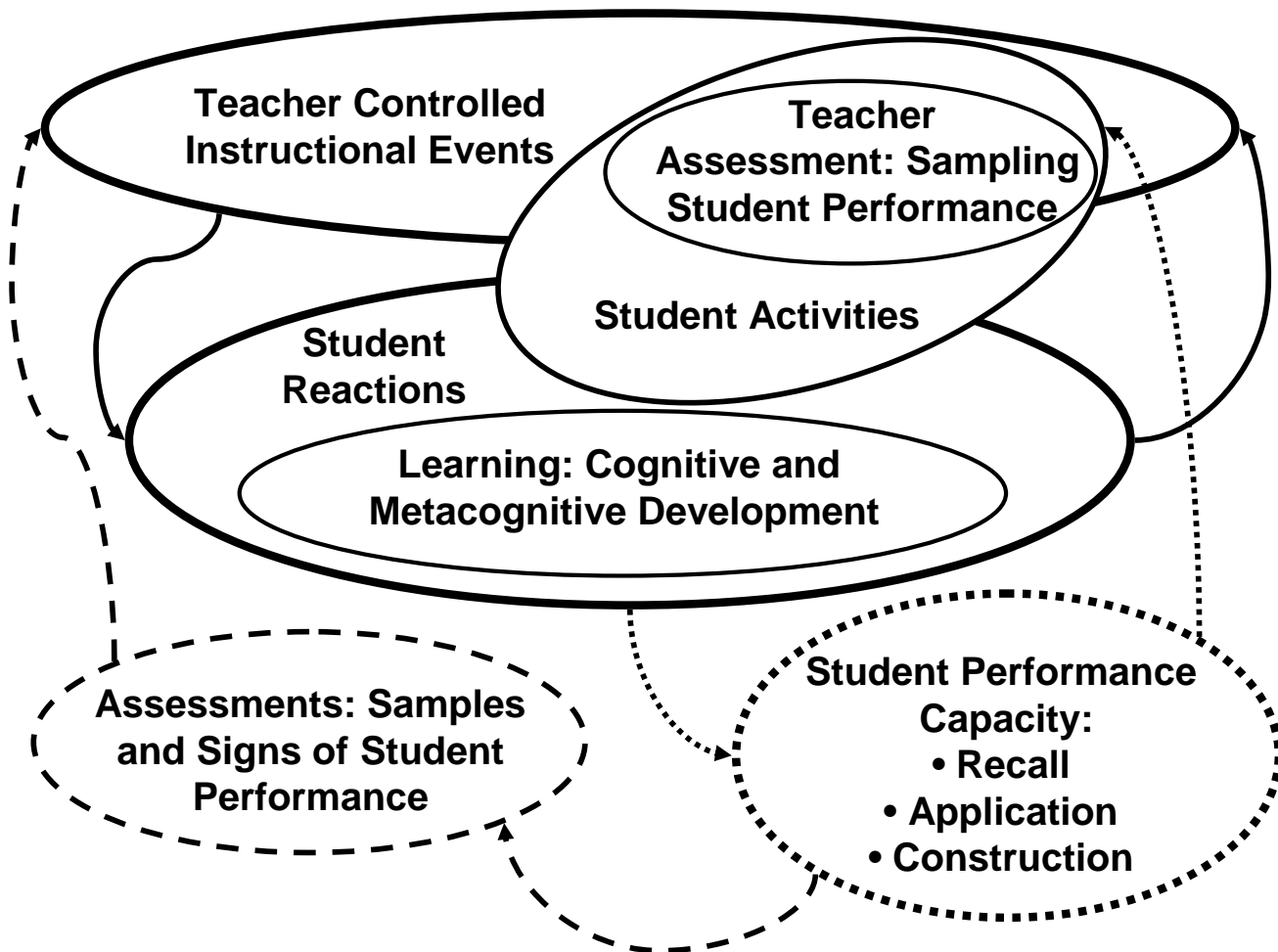
Levels of Aggregation

- ◆ Different issues apply to different levels of aggregation. These aggregations include:
 - Student achievement scores
 - School yearly assessment indexes
 - District yearly assessment indexes
- ◆ ...with time as an additional factor
 - School accountability cycle results
 - District accountability cycle results
 - Longitudinal student growth – across and within years

There is a Much Needed Role for Model Building

- ◆ ... and not just psychometric models.
- ◆ What are the assumptions that link, on organizational, social-cultural, and psychological levels, assessment test score use to changes in learning outcomes?
- ◆ How is test information to be used in attempting to achieve educational goals?
 - Information about test content and format which is used prior to test administration, including information about expected student performance.
 - Test scores released after test administration.
 - Individual level
 - Aggregated to teacher level
 - Aggregated to school level
 - Aggregated to district level
- ◆ Where are the theories that can help us build tenable assumptions?
.....multi-disciplinary thoughts are needed

Sample – Just for fun



Instruction, learning, performance, and assessment.

Advertisement

- ◆ “Google”
 - Kentucky Department of Education
 - Pennsylvania Board of Education
 - California Department of Education
 - Idaho Board of Education
 - NAEP

- ◆ and search for:
 - HumRROor
 - Human Resources Research Organization