Constructing Valid Performance Assessments –
The View from the Shoulders of the Giants

www.rasch.org/memos.htm

“If I have seen further, it is only by standing on the shoulders of giants.”
Isaac Newton, 1676

From a portrait by Kneller in 1689
These are questions of validity...

“Can we give oral examinations effectively? ... Do we wish to measure speaking ability and do we know what speaking ability is? ... These are questions of validity and administration.”


www.joytalk.co.jp/ladosys/index.html
Validity is “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.”

In 1998, Sam Messick agreed to speak at LTRC, but he died before that happened. In his honor, ETS sponsor these lectures:


2000 **Merrill Swain**, Examining Dialogue: Another Approach to Validating Inferences Drawn from Test Scores?

2001 **Richard Luecht**, New Directions in Computerized Testing Research


2003 **Patricia Broadfoot**, Dark Alleys and Blind Bends: Testing the Language of Learning

2004 **Eva Baker**, Language, Learning, and Assessment: Improving Validity
Samuel J. Messick
Memorial Award Lectures

2005 Bruno Zumbo, Reflections on Validity at the Intersection of Psychometrics, Scaling, Philosophy of Inquiry, and Language Testing

2006 Mark Wilson, Building out the Measurement Model to Incorporate Complexities in Language Testing

2007 Robert J. Mislevy, Toward a Test Theory for the Interactionalist Era

2008 James Dean Brown, Why don’t the Stakeholders in Language Assessment Just Cooperate?

2009 Lorrie A. Shepherd, Understanding Learning (and Teaching) Progressions as a Framework for Language Testing

2010 Michael Kane, Validating Score Interpretations and Uses

2011 John “Mike” Linacre, Constructing Valid Performance Assessments – the View from the Shoulders of the Giants
25 years ago, a judging problem …

Mary Lunz at ASCP:

- Certifying medical technicians
- 15 slides of human body-parts for each technician
- Expert, expensive raters (pathologists)
- Each slide only viewed once by only one rater

http://nmnwse.org/careers/HTML/C33HISTO.HTM
The biggest threat to validity …

“I find the element of chance in these public examinations to be such that only a fraction - from a third to two-thirds - of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed.”

Edgeworth, 1890, The Element of Chance in Competitive Examinations, JRSS
“Inconsistencies among raters will create problems in generalizing .... Conclusions about [the candidates] would depend on the luck of the draw - a “liberal” rater rather than a “stringent one”.

Shavelson & Webb, Generalizability Theory – A Primer, 1991, p. 8-9
… or a Measurement Challenge?

Judges are like weight scales. We must adjust for their zero-calibration in order to obtain accurate measures.

The measures are superficial, instantaneous and fleeting. We need to quantify an amount now. We need the measure to have the meaning we want it to have now.
1. Decide in advance:
What is a valid rater?

a) Valid raters give the “correct” rating: rating machines. Rater training helpful, but ... what are the correct ratings?

b) Valid raters agree with each other on the exact rating. Paired raters with third for disagreements.

c) Valid raters agree on which performances are better and which are worse = high correlation.

Trained raters, part of an ETS study.
Agreement: Exact ≠ Ordered

A comparative study of paired judges:

** most disagreements, higher correlation

* 10 or more disagreements

Lower disagreements, lower correlation

A Preliminary Study of Raters for the Test of Spoken English. Isaac Bejar, 1985. ETS: RR-85-05
Dependable Independent Experts

1. They agree on overall competence = expert (b.)

2. They disagree on details = independence (c.)

3. They maintain their own standards of leniency and severity = dependable.
   
   *Train for stability, not for agreement!*
2. Design for Comparability

California students are rated higher than Michigan students – but why?

California students better?

California raters lenient?

Must have linked networks of judges – especially in paired-judge designs!
The Minimum-Effort Judging Plan

A simplified version of the ASCP judging plan:
Each Essay by each Person rated once by one different Judge
3. Begin Analyzing when Data Collection Starts

*International scandal:* French judge influenced by the Russian Mafia!

*Problem discovered too late* 😞

www.guardian.co.uk/world/2002/aug/01/russia.sport

Olympic ice-skating: Pairs Skating:
Winter Olympics, Salt Lake City 2002
If analysis had started promptly …

French judge’s misbehavior detected after the Short Program

Not too late to obtain valid final scores from the later Free Program
4. Additive Measures

“One more” means the same amount extra, no matter

www.rasch.org/memo62.htm
5. When in doubt, what would a physical scientist do?

Two measurement crises:

1890 – Francis Ysidro Edgeworth - variation of ratings

1887 – the Michelson-Morley experiment - constancy of the speed of light

Two different reactions ...

www.epola.co.uk/epola_org/michelson.htm
Physicists: Size, not Significance

“It appears, from all that precedes, reasonably certain that if there be any relative motion between the earth and the luminiferous ether, it must be small;”

= speed of light is constant in a vacuum


www.rasch.org/rmt/rmt111c.htm
Medical Researchers (Social Scientists?): Significance, not Size

Most medical studies are wrong

By Dana Blankenhorn | October 20, 2010, 6:26 AM PDT

"Science is a noble endeavor, but it's also a low-yield endeavor," Dr. John Ioannidis told The Atlantic recently.

It may be the truest statement yet made on medical research.

It's a story flying around the medical community today, although it's based on a single five-year old study, from a team of Greek researchers headed by Ioannidis, titled simply Why Most Published Research Findings are False.

[Link to the article]
Messick: Theory + Data \( \rightarrow \) Validity

_A desperate email from E.L. received in June, 2011:_

Question: I understand the **importance of having a substantive theory** of measurement in social sciences, but with **absence of such a theory** will simulation gives us a little help to find a better solution? Otherwise, **I don't know what we are working for** and where we're working towards, if we don't even know whether or not the method we choose is improved or say better than some others....
From the shoulders of the giants, we can see a bright horizon …

theory construction

In 1911, Albert Einstein predicted that gravity would bend light.

April 2011: gravity bending light: Lensing Cluster Abell 383

Albert Einstein, 1921, Nobel Prize picture. [image]

Perhaps 2011 is the “Breakthrough” year for Social Science Theory!

Linacre’s Messick Memorial Lecture is accessible at:

www.rasch.org/memos.htm