Samuel J. Messick Memorial Lecture - LTRC, Michigan, June 2011

John Michael Linacre

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

www.rasch.org/memos.htm

From a portrait by Kneller in 1689

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

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."

Robert Lado (1915-1995)

Robert Lado (1960) "English Language Testing: Problems of Validity and Administration", ELT Journal, 14, 4, 153-161.

www.joytalk.co.jp/ladosys/index.html

… empirical evidence and theoretical rationales …

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."

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Washington, DC: American Council on Education / Macmillan.

Samuel J. Messick (1931-1998)

edweek.org

Samuel J. Messick Memorial Award Lectures

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

1999 Tim McNamara, Validity in Language Testing: The Challenge of Sam Messick's Legacy.

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

2001 Richard Luecht, New Directions in Computerized Testing Research

2002 Geofferey Masters, 20th Century Foundations for 21st Century Measurement

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

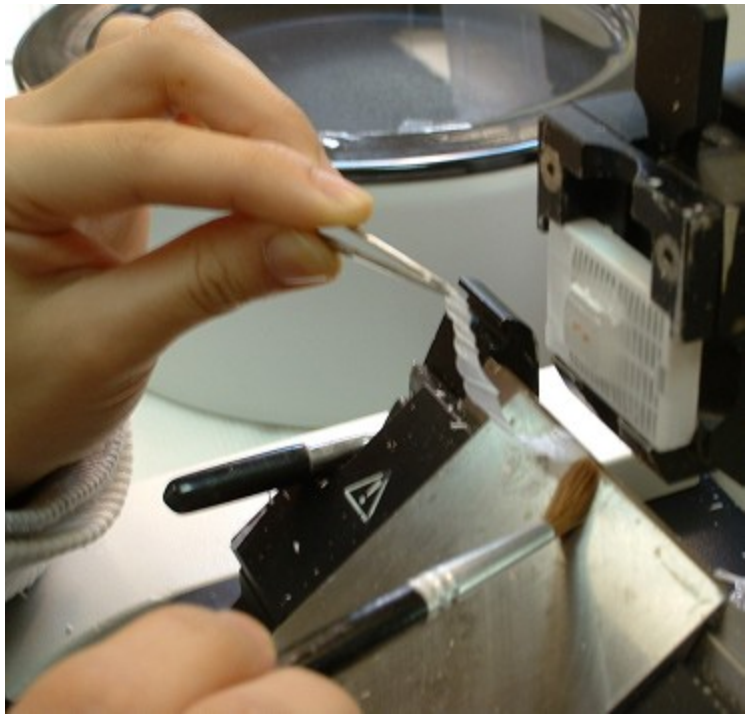
2010 Michael Kane, Validating Score Interpretations and Uses

2011 John "Mike" Linacre, Constructing Valid Performance Assessments – the View from the Shoulders of the Giants

25 years ago, a judging problem …

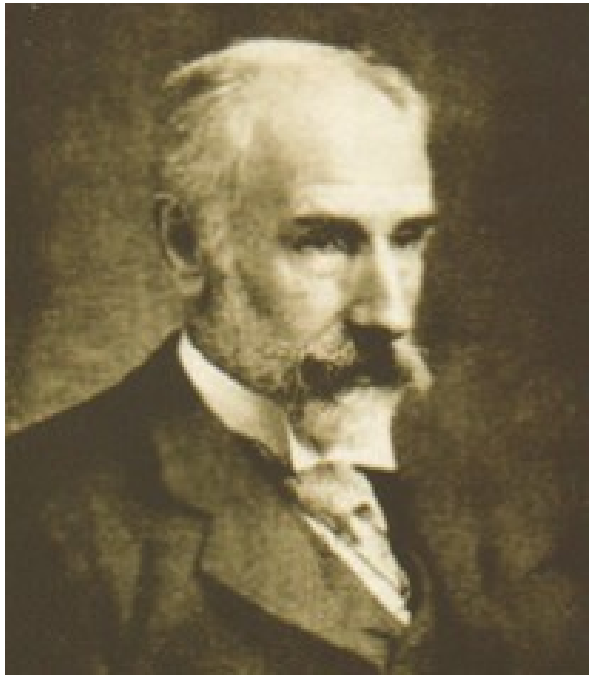
Mary Lunz at ASCP:

- 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 …



Francis Ysidro Edgeworth (1845-1926)

nowosci.bg.univ.gda.pl

"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*

The Luck of the Draw …

"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, Generalizabilty Theory – A Primer, 1991, p. 8-9

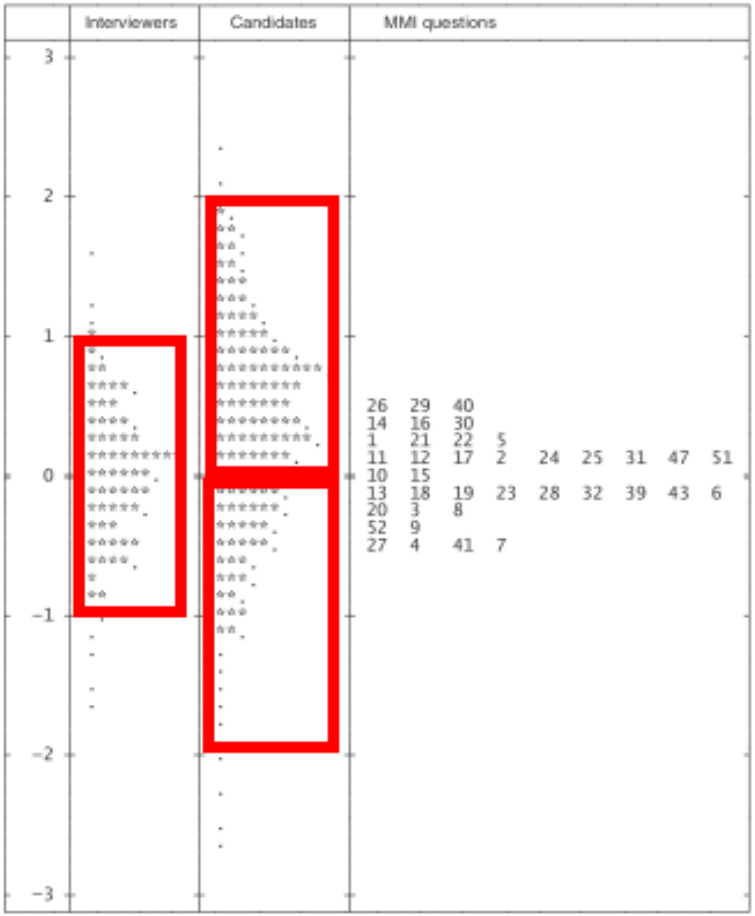


Table 6 as published in "Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview?" by Chris Roberts, Imogene Rothnie, Nathan Zoanetti & Jim Crossley, Medical Education 2010: 44: 690–698

… or a Measurement Challenge?

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

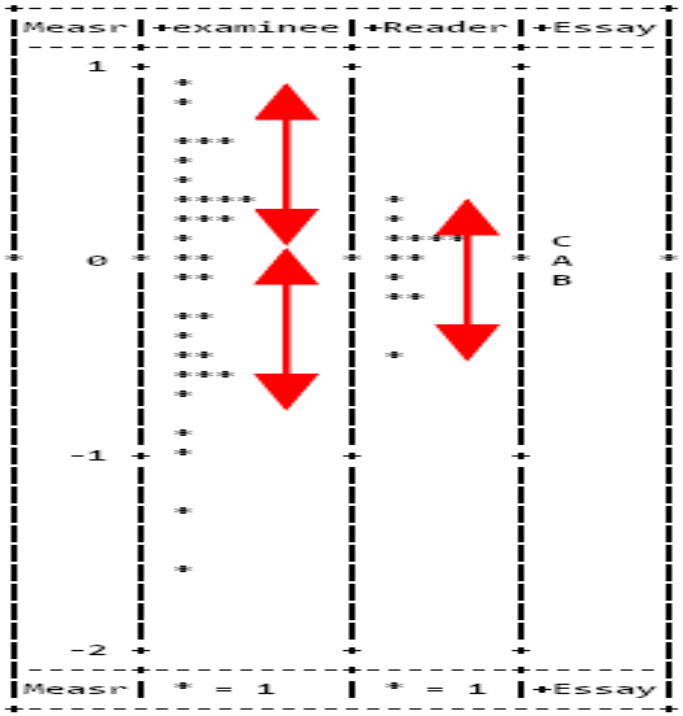
The **measures are superficial**, instantaneous and fleeting. We need to quantify an amount now. We need the measure to have the meaning we want it to have now.

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. Henry I. Braun. Understanding score reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, Spring 1988, 13/1, 1-18

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

Correlations of Individual Raters with Paired Raters on Each Linguistic Dimension

Rater ID.	N	PRON	GRAM	FLU	COMP	
111	93	.75	.77	.61	.81	*
113	141	.72	.81	.62	.68	
114	59	.87	.75	.74	.83	*
118	174	.77	.80	.71	.76	
120	151	.79	.85	.75	.82	**
121	119	.82	.82	.70	.77	*
124	39	.83	.89	.82	.88	
125	13	.74	.87	.88	.90	
126	22	.77	.93	.88	.89	
127	33	.89	.93	.79	.88	
128	75	.84	.91	.86	.92	
129	89	.80	.85	.81	.82	*
130	13	.23	.80	.57	.82	
135	75	.87	.90	.87	.91	

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!***

www.doodysfancydress.co.uk

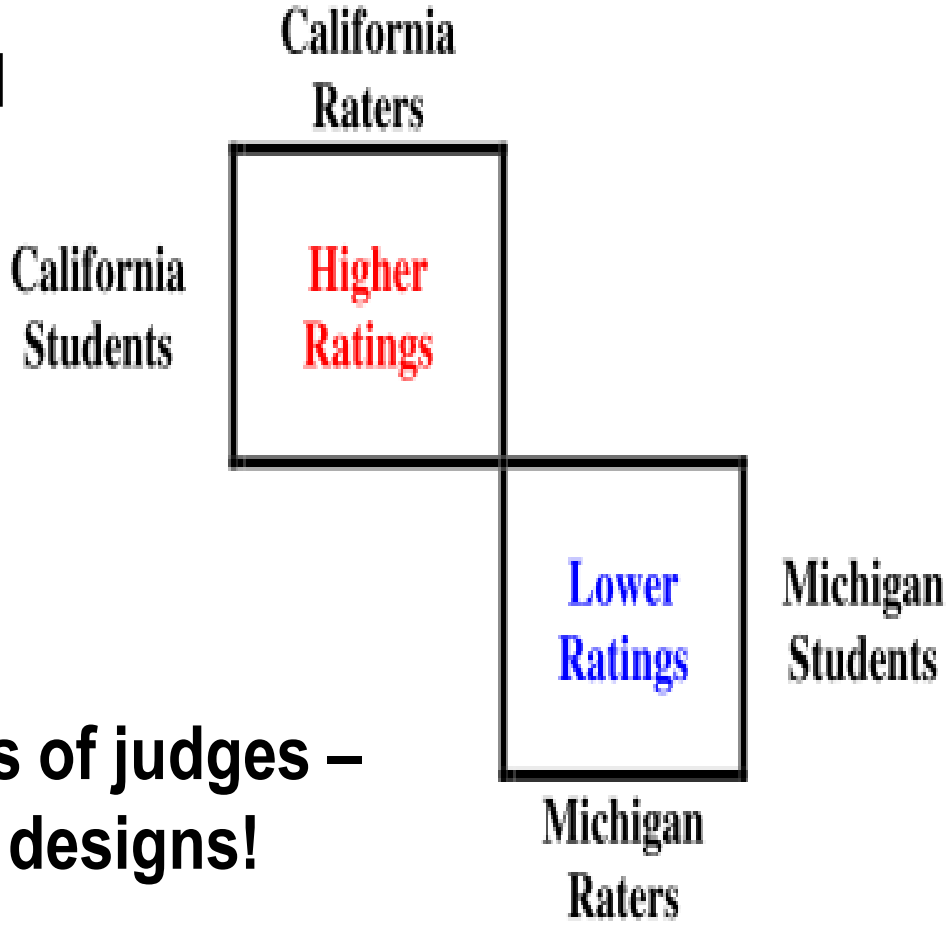
2. Design for Comparability

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

Judge	1	2	3	4	5	6	7	8	9	10	11	12
Essay	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC	ABC
Person												
1							6	7	5			
2				5	3					5		
3	4									5	3	
4		2				5				4		
5					7					4	5	
6	5	6									6	
7					5		3					4
8		6						6		5		
9					3		6	4				
10				7	5				7			
11									6		4	3
12	4			6								7
13	4			6					4			
14	6	3					4					
15	4								4	6		
16		6	4								8	
17			2	6	6							
18						4	4					4
19		7		6		4						

A simplified version of the ASCP judging plan:

Each Essay by each Person

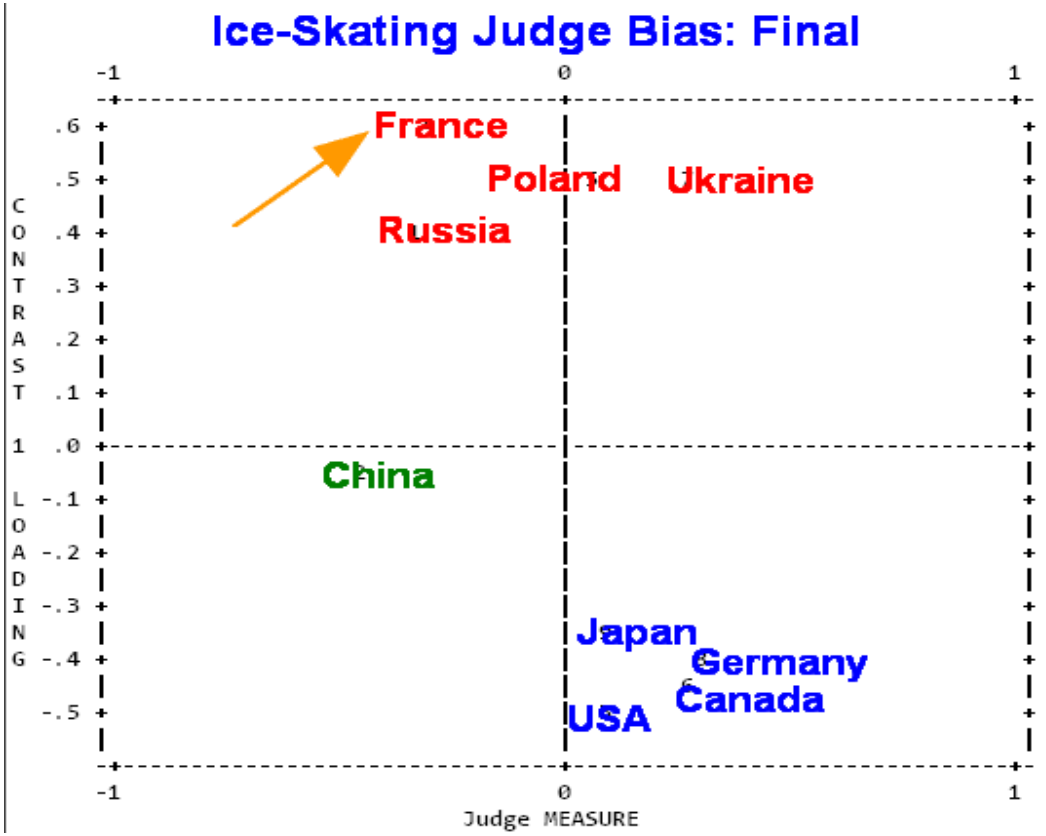
rated once by one different Judge

3. Begin Analyzing when Data Collection Starts

International scandal:

French judge influenced by the Russian Mafia!

Problem discovered too late



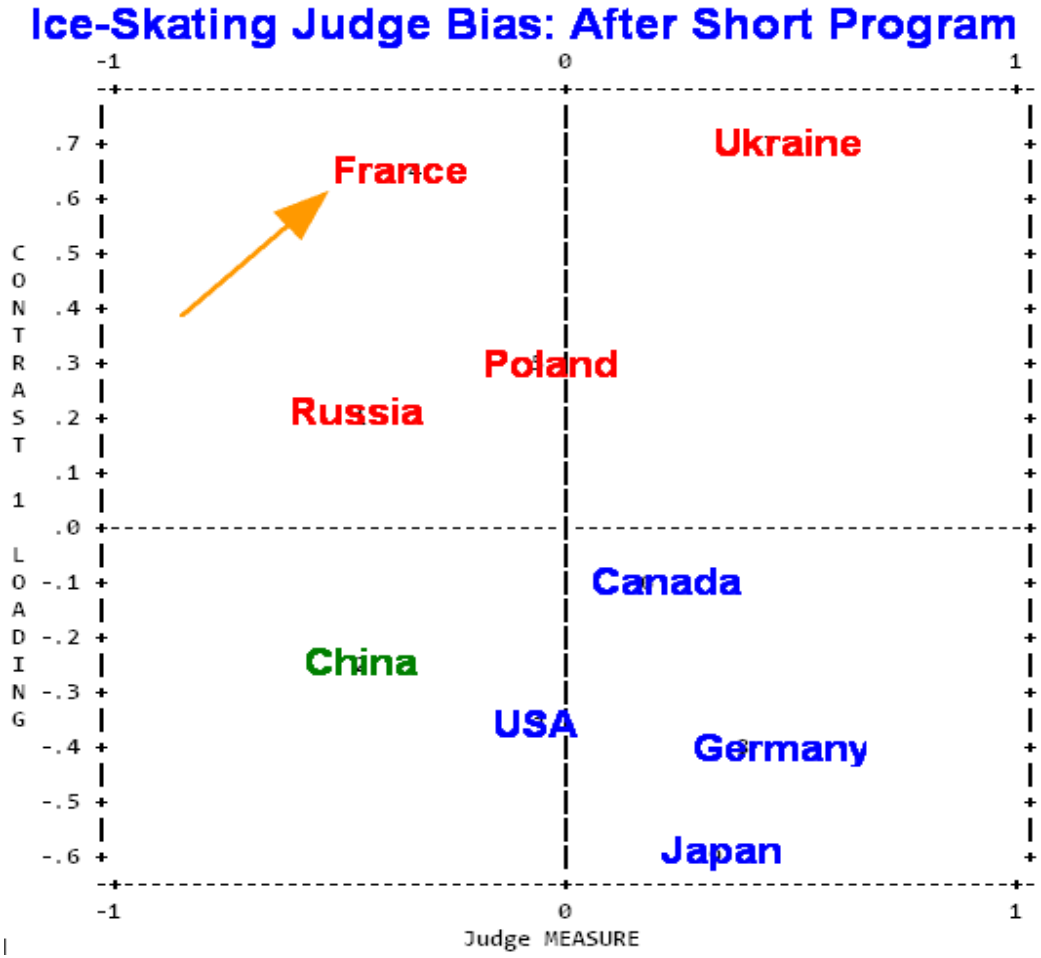
www.guardian.co.uk/world/2002/aug/01/russia.sport

Olympic ice-skating: Pairs Skating:
Winter Olympics, Salt Lake City 2002

If analysis had started promptly …

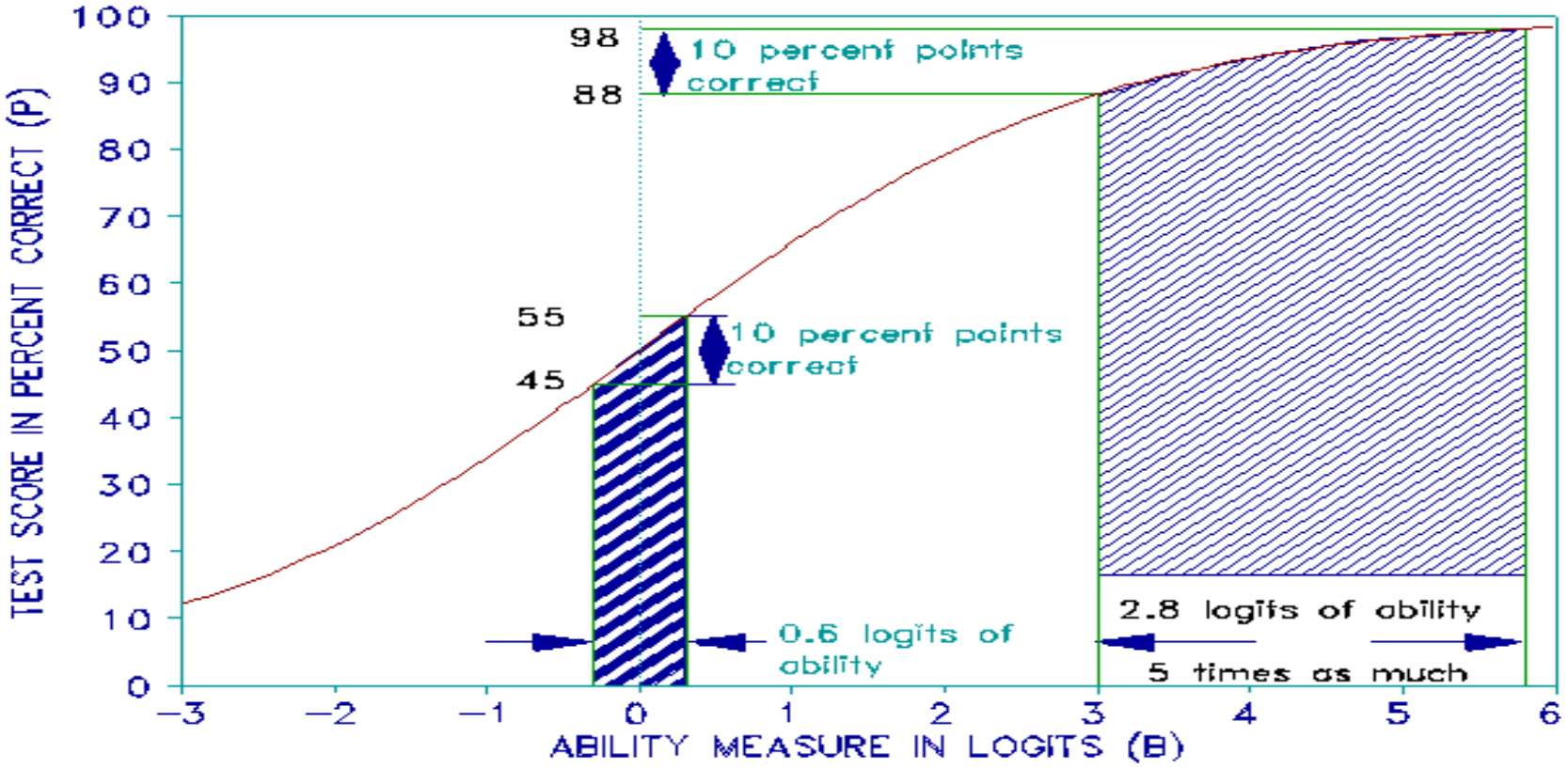
French judge's **misbehavior detected** after the Short Program

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 …

A.A. Michelson
1852 - 1931

E.W. Morley
1838 - 1923

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

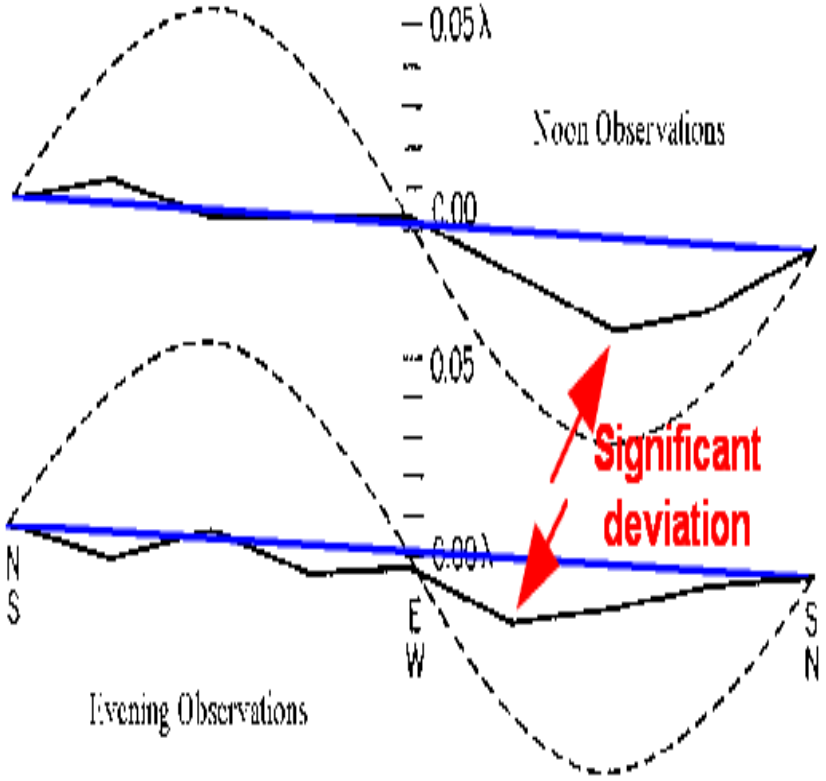


Figure 6 from Mickelson and Morley (1887)

Michelson, Albert Abraham & Morley, Edward Williams (1887). "On the Relative Motion of the Earth and the Luminiferous Ether". American Journal of Science 34: 333–345.

www.rasch.org/rmt/rmt111c.htm

Medical Researchers (Social Scientists?): Significance, not Size

Home / Technology / Rethinking Healthcare

Follow this blog:

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.

www.smartplanet.com/blog/rethinking-healthcare/most-medical-studies-are-wrong/1763

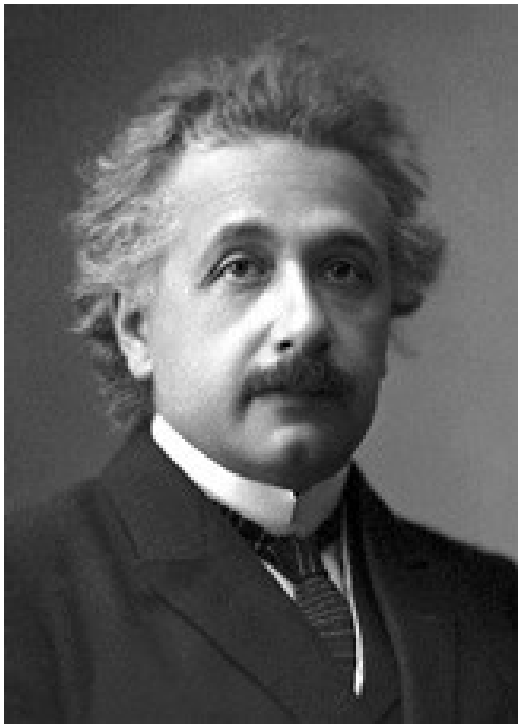
Messick: Theory + Data →Validity

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



Albert Einstein, 1921, Nobel Prize picture.

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

April 2011: **gravity bending light**: Lensing Cluster Abell 383

.eurekalert.org/pub_releases/2011-04/eic-fgw041111.php

Perhaps 2011 is the "Breakthrough" year for **Social Science Theory**!

Linacre's Messick Memorial Lecture is accessible at:

www.rasch.org/memos.htm