Methodological and Statistical Issues in Research Proposals

Measures of all things

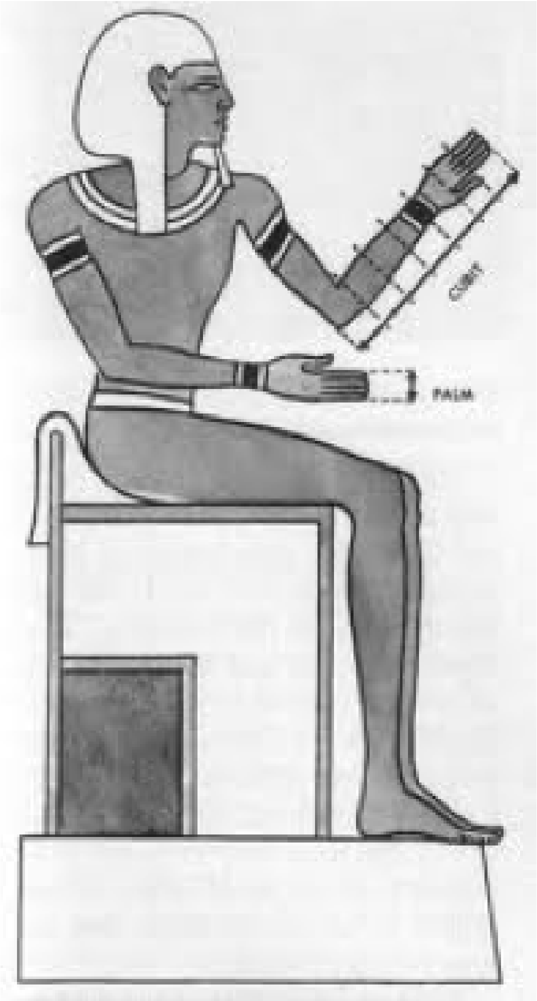
Reliability and validity of research measures and how these impact your research proposals

Rich Jones

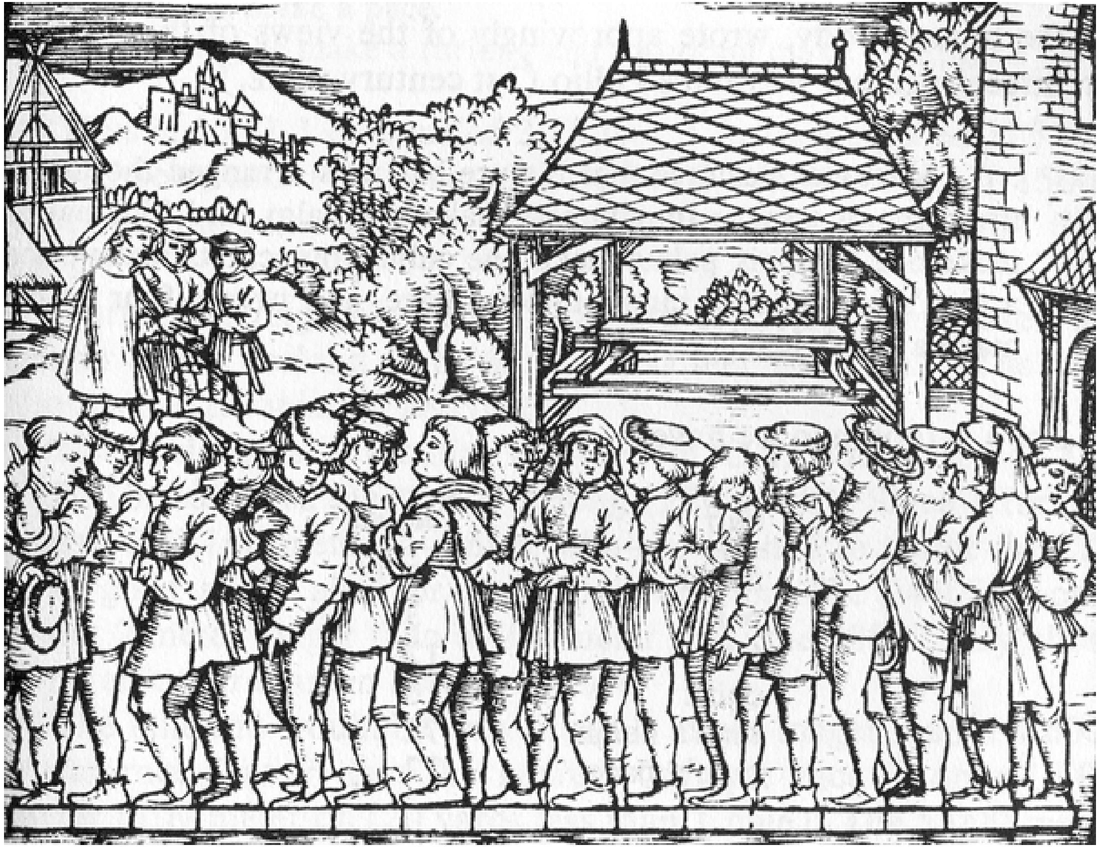
NIDUS/CEDARTREE, 7th Annual Delirium Boot Camp

October 28, 2019, Penn Stater, State College PA

rich_jones@brown.edu 🐦 @rnjma



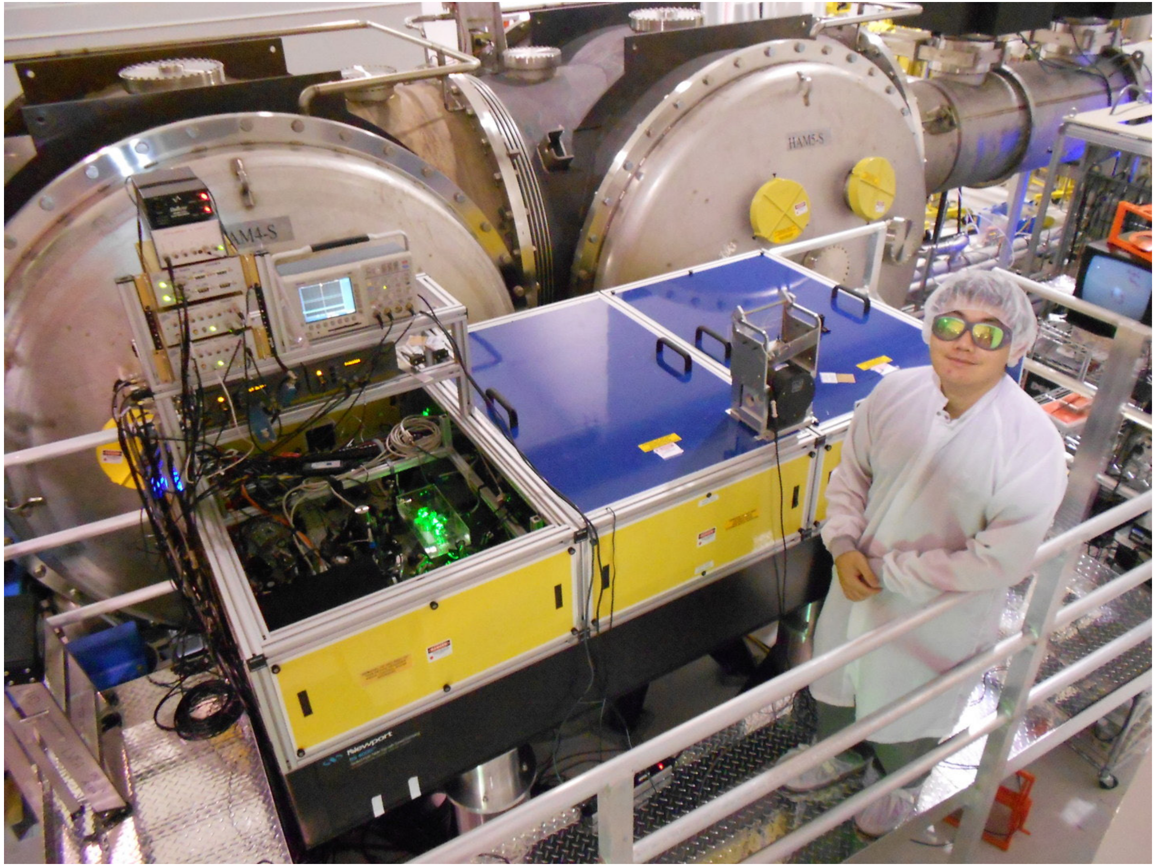
http://abemkemet.blogspot.com/2012/06/ruler-for-ruler.html



https://en.wikipedia.org/wiki/Foot_(unit)

Jakob Köbel (1460 - 1533) - Geometrei. Von künstlichem Feldmessen und absehen (published first in **1535 or 1536,** reprinted in 1608.[1])

Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be the right and lawful foot.



https://www.ligo.org/science/Publication-SqueezedVacuum/index.php

1983

The meter is defined as the length of the path travelled by light in a vacuum in 1/299,792,458 of a second

https://en.wikipedia.org/wiki/Metre

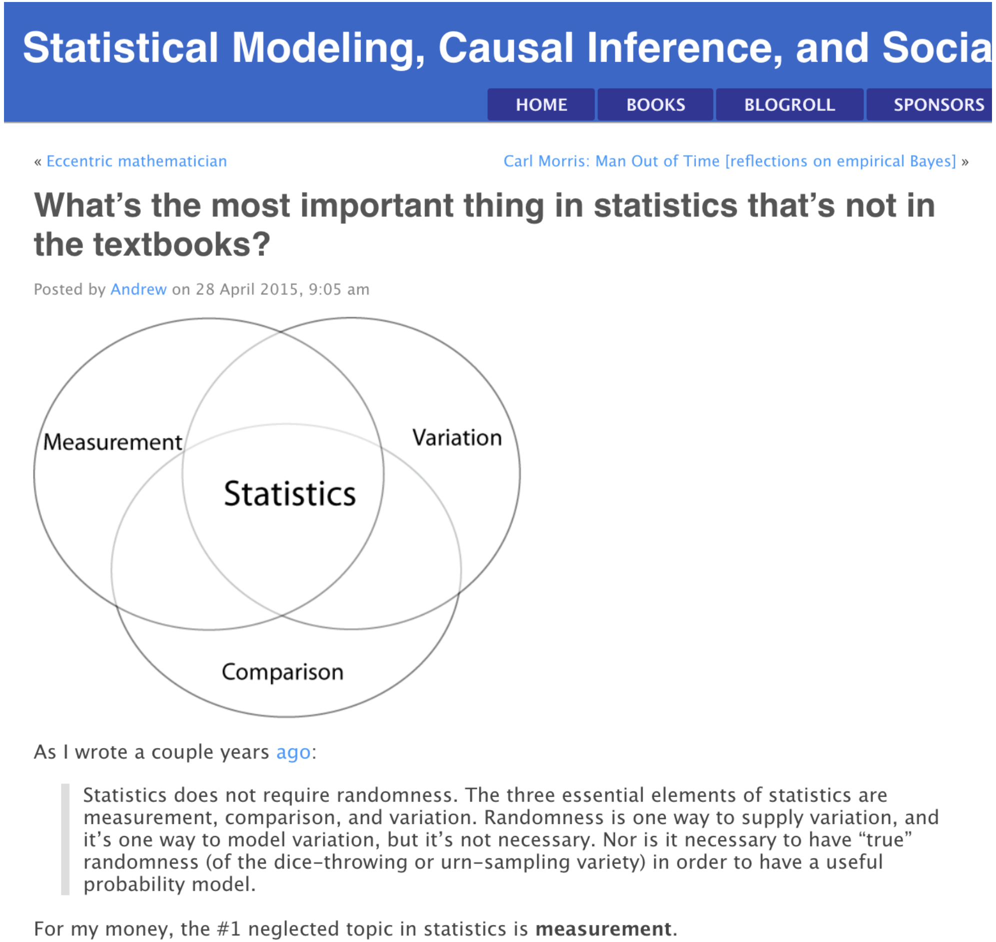
c = speed of light in a vacuum

c = 299,792,458 m/s

Measurements are developed to address a specific practical need.

Measurements are refined as the need arises (e.g., for greater precision arises), often in the context of some new use.

Refinement of measures is linked with technological development.



« Eccentric mathematician

Carl Morris: Man Out of Time [reflections on empirical Bayes] »

What’s the most important thing in statistics that’s not in the textbooks?

Posted by Andrew on 28 April 2015, 9:05 am

As I wrote a couple years ago:

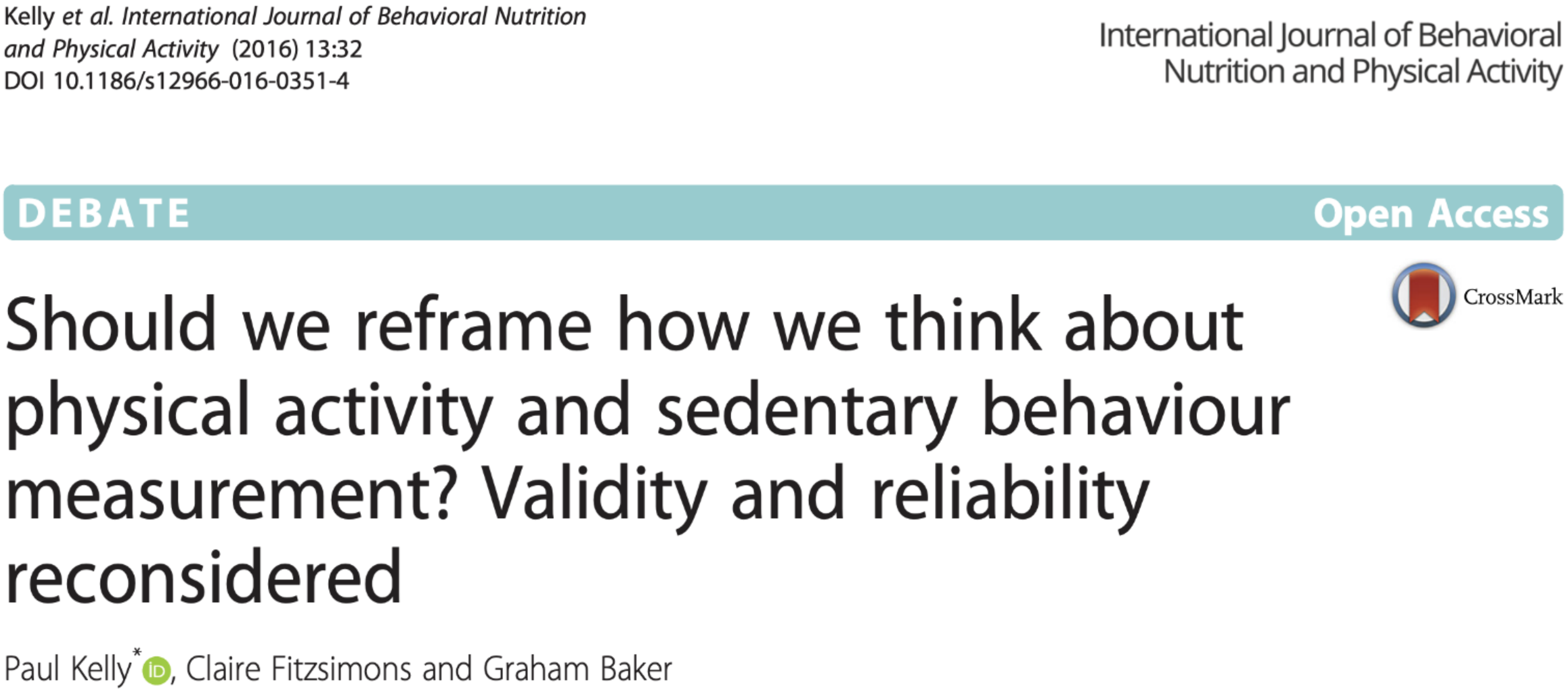
> Statistics does not require randomness. The three essential elements of statistics are measurement, comparison, and variation. Randomness is one way to supply variation, and it’s one way to model variation, but it’s not necessary. Nor is it necessary to have “true” randomness (of the dice-throwing or urn-sampling variety) in order to have a useful probability model.

For my money, the #1 neglected topic in statistics is **measurement**.

The #1 neglected topic in statistics is measurement

- Andrew Gelman

https://andrewgelman.com/2015/04/28/whats-important-thing-statistics-thats-not-textbooks/



Kelly *et al. International Journal of Behavioral Nutrition and Physical Activity* (2016) 13:32
DOI 10.1186/s12966-016-0351-4

International Journal of Behavioral Nutrition and Physical Activity

DEBATE

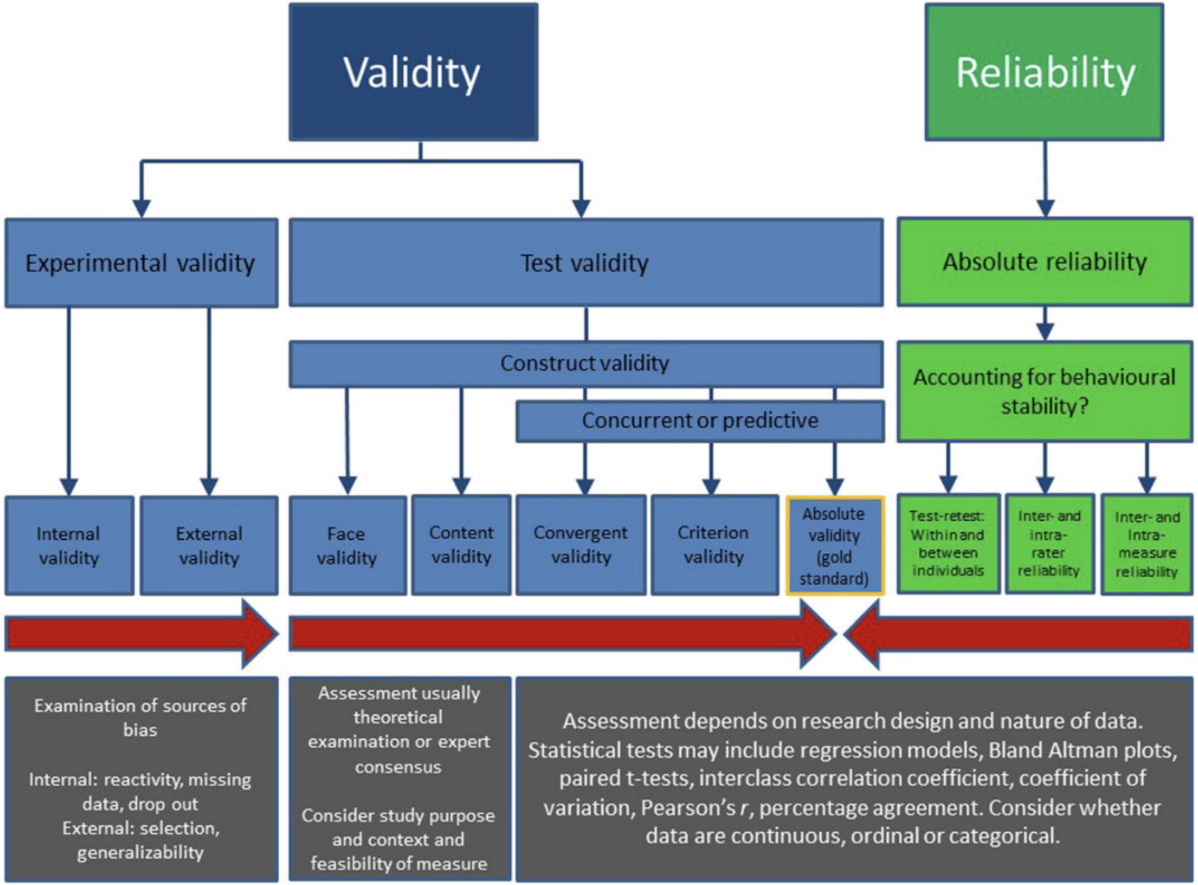
Open Access

Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered

Paul Kelly*, Claire Fitzsimons and Graham Baker

Terminology is used randomly, synonymously, possibly incorrectly and we all get confused

Already we have used terms that you may have taken issue with. In many places we could have used different terms such as precision, concordance, uncertainty, or accuracy. There are also many sub-types of validity and reliability, some of which we have not yet discussed. For example, construct, comparative, absolute, relative, predictive, discriminant, representation, and translation validity; and inter-rater, intra-rater, relative, or absolute reliability.



Examination of sources of bias

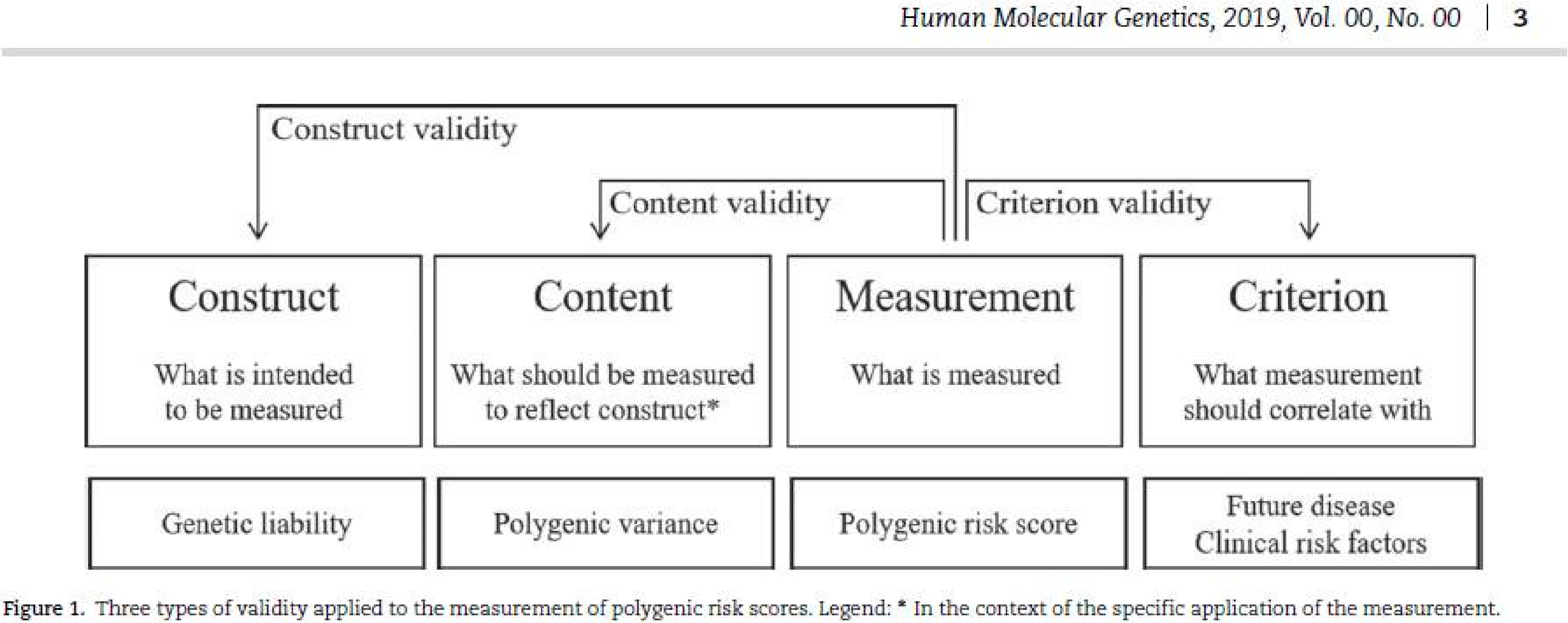
Internal: reactivity, missing data, drop out
External: selection, generalizability

Assessment usually theoretical examination or expert consensus

Consider study purpose and context and feasibility of measure

Assessment depends on research design and nature of data. Statistical tests may include regression models, Bland Altman plots, paired t-tests, interclass correlation coefficient, coefficient of variation, Pearson's *r*, percentage agreement. Consider whether data are continuous, ordinal or categorical.

Fig. 3 The Edinburgh Framework v1.0 for validity and reliability in PA and SB measurement



Human Molecular Genetics, 2019, Vol. 00, No. 00 | 3

Construct validity

Content validity

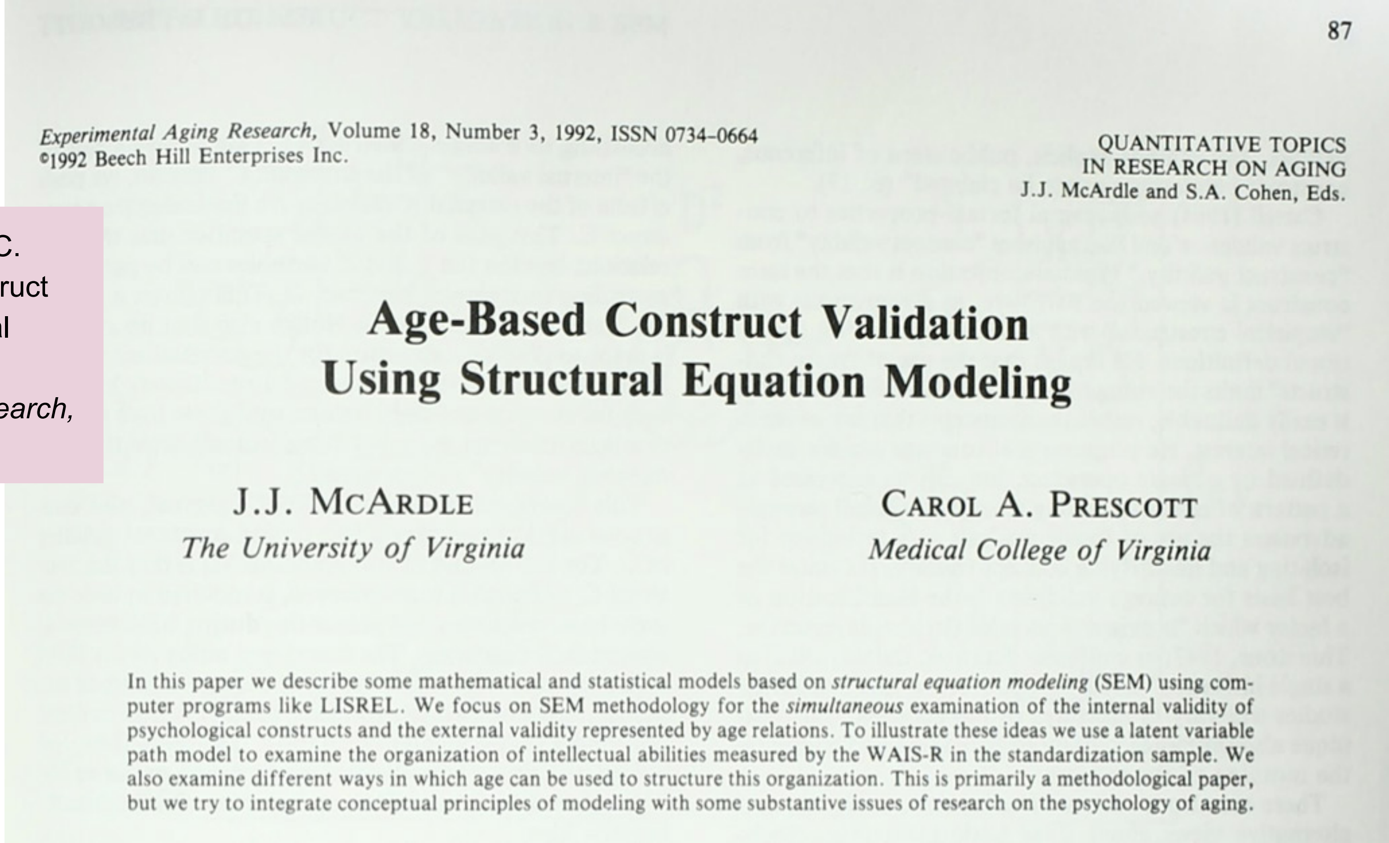
Criterion validity

Construct	Content	Measurement	Criterion
What is intended to be measured	What should be measured to reflect construct*	What is measured	What measurement should correlate with
Genetic liability	Polygenic variance	Polygenic risk score	Future disease Clinical risk factors

Figure 1. Three types of validity applied to the measurement of polygenic risk scores. Legend: * In the context of the specific application of the measurement.

Janssens A. 2019 Human Molecular Genetics, https://doi.org/10.1093/hmg/ddz205

McArdle, J., & Prescott, C. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18*(3), 87-116.



Experimental Aging Research, Volume 18, Number 3, 1992, ISSN 0734-0664
©1992 Beech Hill Enterprises Inc.

QUANTITATIVE TOPICS
IN RESEARCH ON AGING
J.J. McArdle and S.A. Cohen, Eds.

Age-Based Construct Validation Using Structural Equation Modeling

J.J. McArdle
The University of Virginia

Carol A. Prescott
Medical College of Virginia

In this paper we describe some mathematical and statistical models based on *structural equation modeling* (SEM) using computer programs like LISREL. We focus on SEM methodology for the *simultaneous* examination of the internal validity of psychological constructs and the external validity represented by age relations. To illustrate these ideas we use a latent variable path model to examine the organization of intellectual abilities measured by the WAIS-R in the standardization sample. We also examine different ways in which age can be used to structure this organization. This is primarily a methodological paper, but we try to integrate conceptual principles of modeling with some substantive issues of research on the psychology of aging.

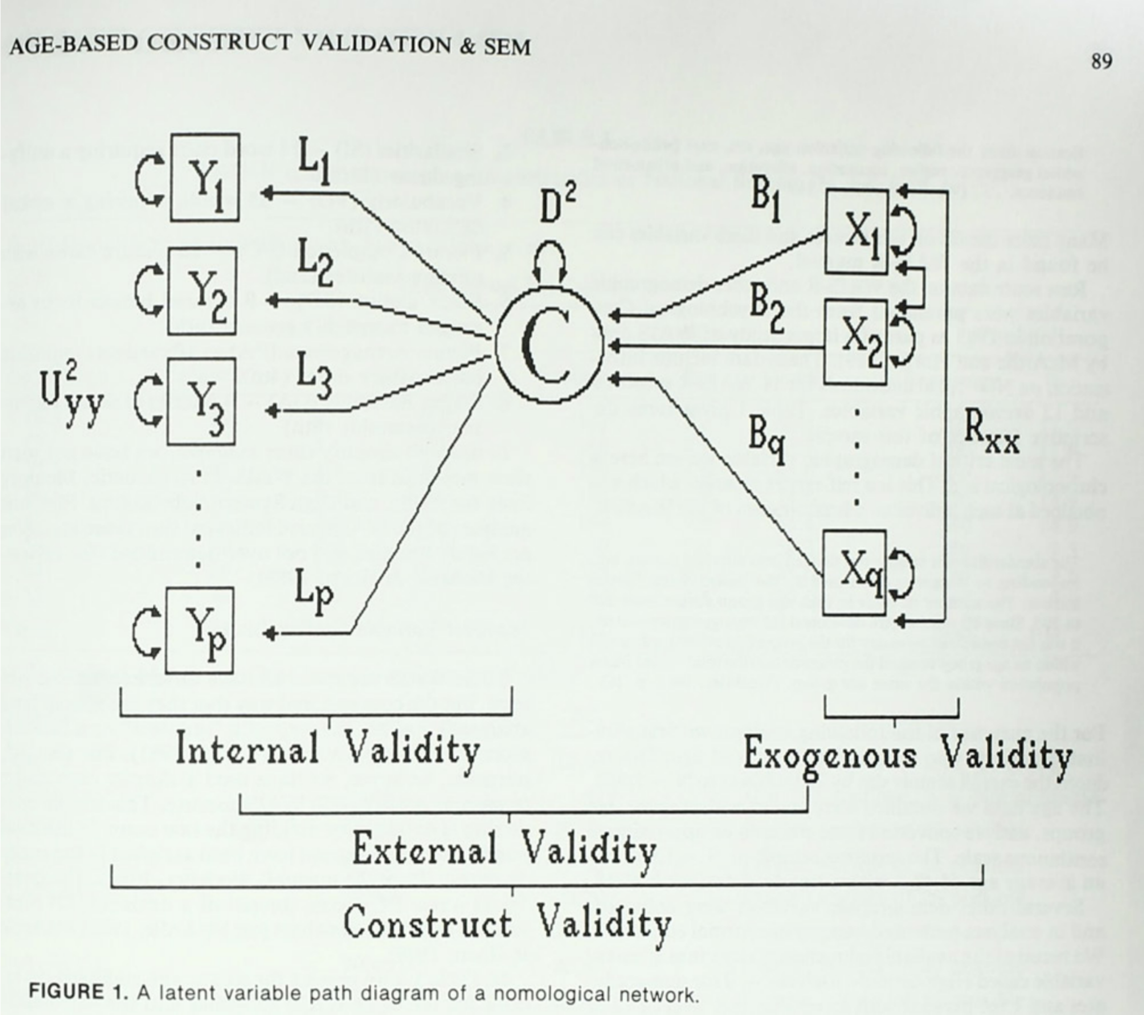
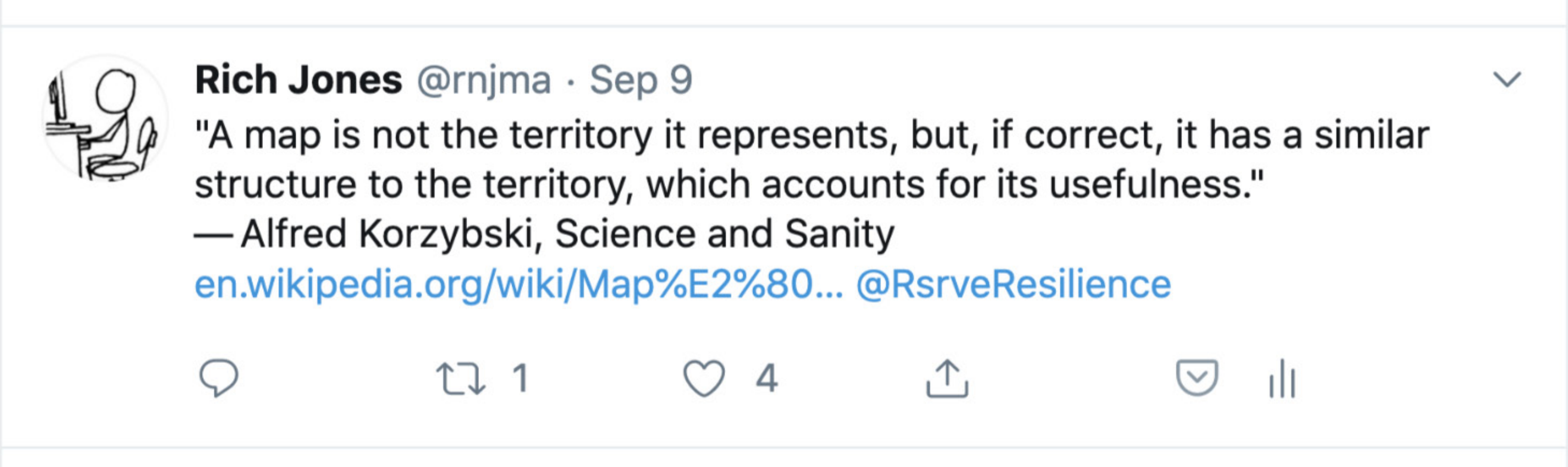


FIGURE 1. A latent variable path diagram of a nomological network.

Reliability	How well do we measure the thing?
Validity	How well do our measurements the thing we measure ***map*** on to the construct we want to measure?



Rich Jones @rnjma · Sep 9

"A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness."
— Alfred Korzybski, Science and Sanity
en.wikipedia.org/wiki/Map%E2%80... @RsrveResilience

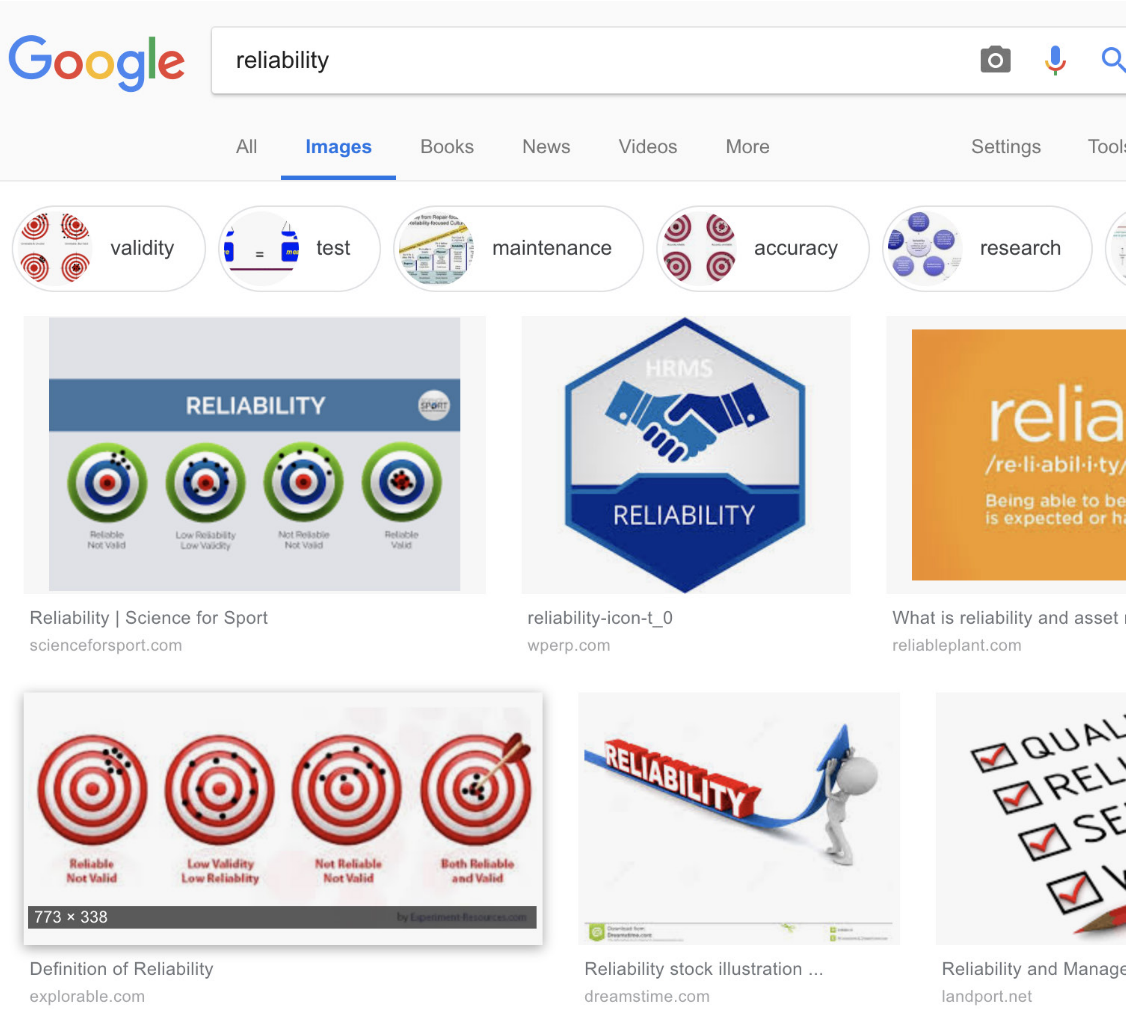
Outline

Reliability ...

1. Concept
2. Basics
3. Implications
4. Paradox
5. Optimizing

https://www.frontiersin.org/articles/10.3389/fnhum.2011.00002/full

Concept: Reliability



reliability

Reliability | Science for Sport
scienceforsport.com

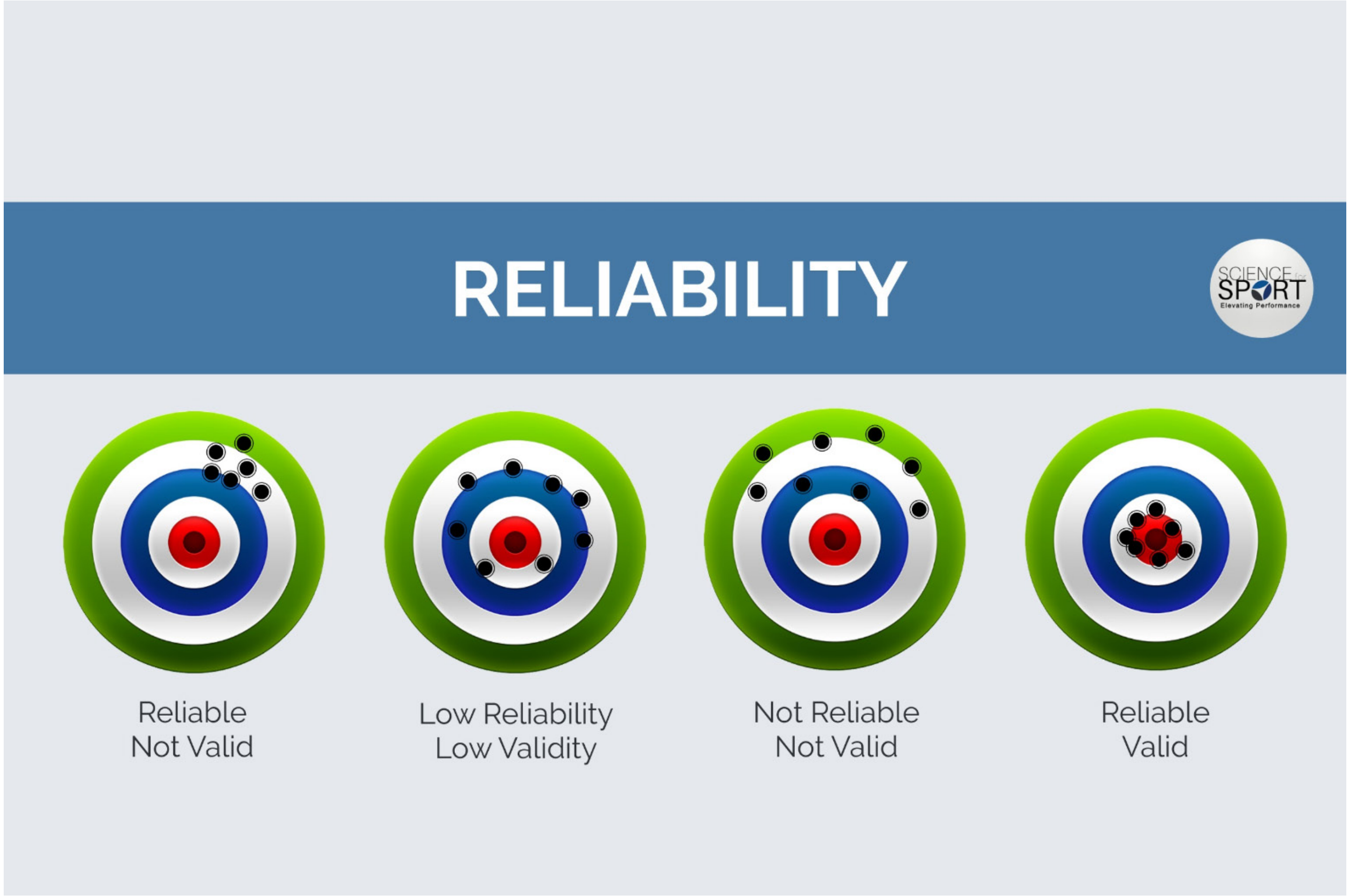
reliability-icon-t_0
wperp.com

What is reliability and asset
reliableplant.com

Definition of Reliability
explorable.com

Reliability stock illustration ...
dreamstime.com

Reliability and Manage
landport.net



RELIABILITY

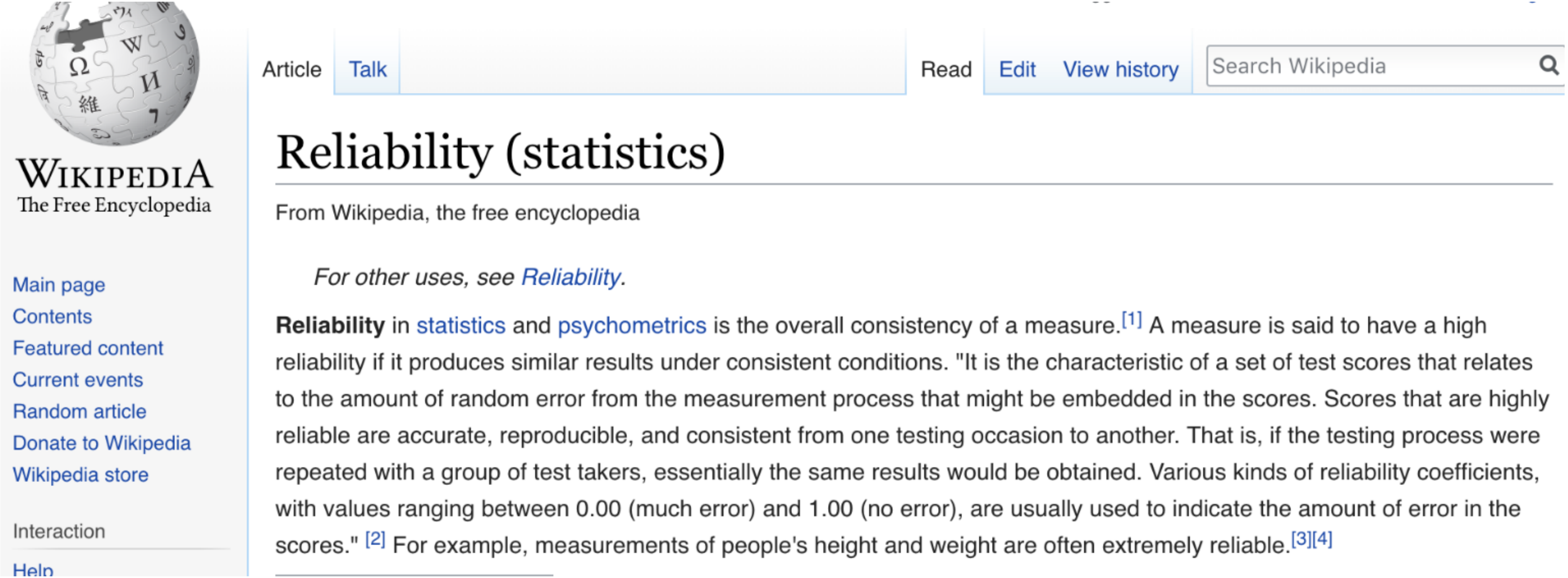
Reliable
Not Valid

Low Reliability
Low Validity

Not Reliable
Not Valid

Reliable
Valid

https://www.scienceforsport.com/reliability/



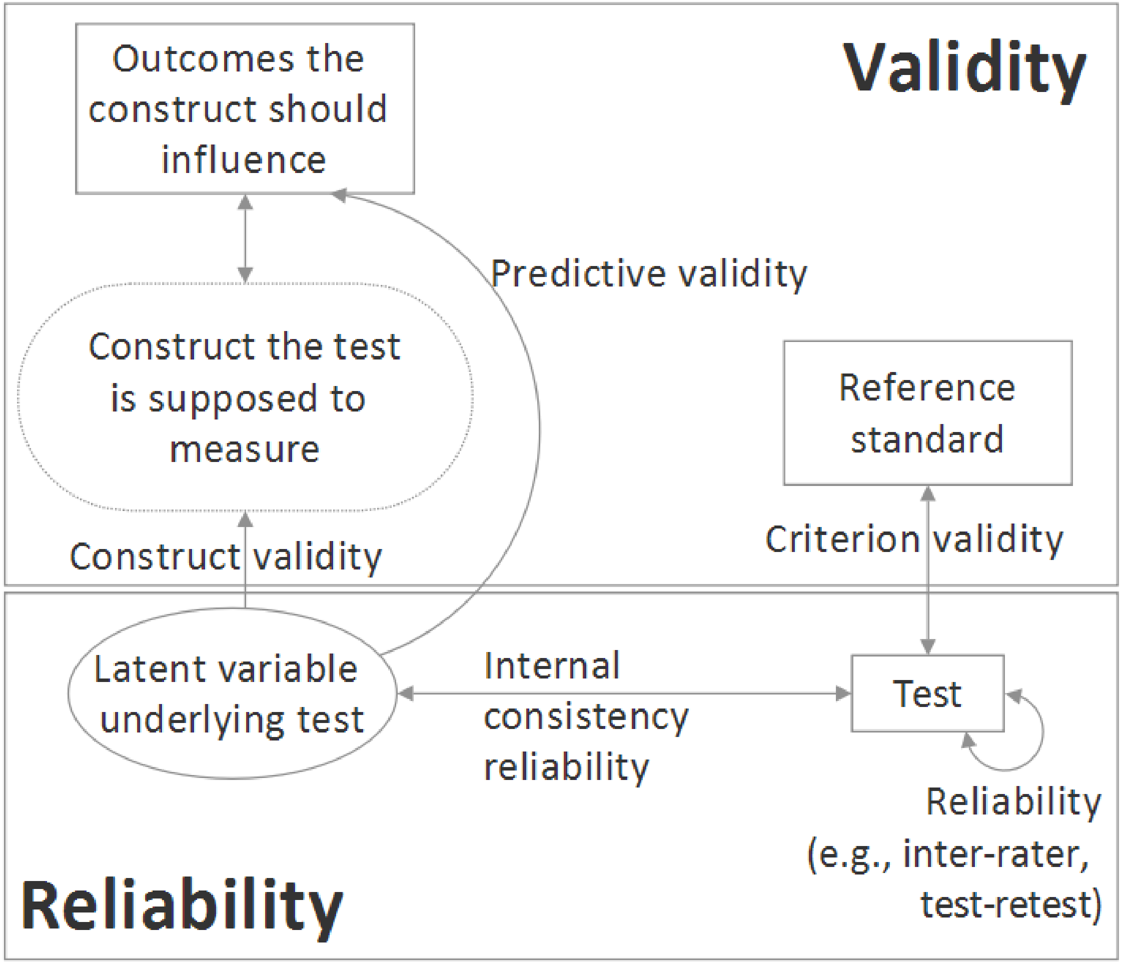
Reliability (statistics)

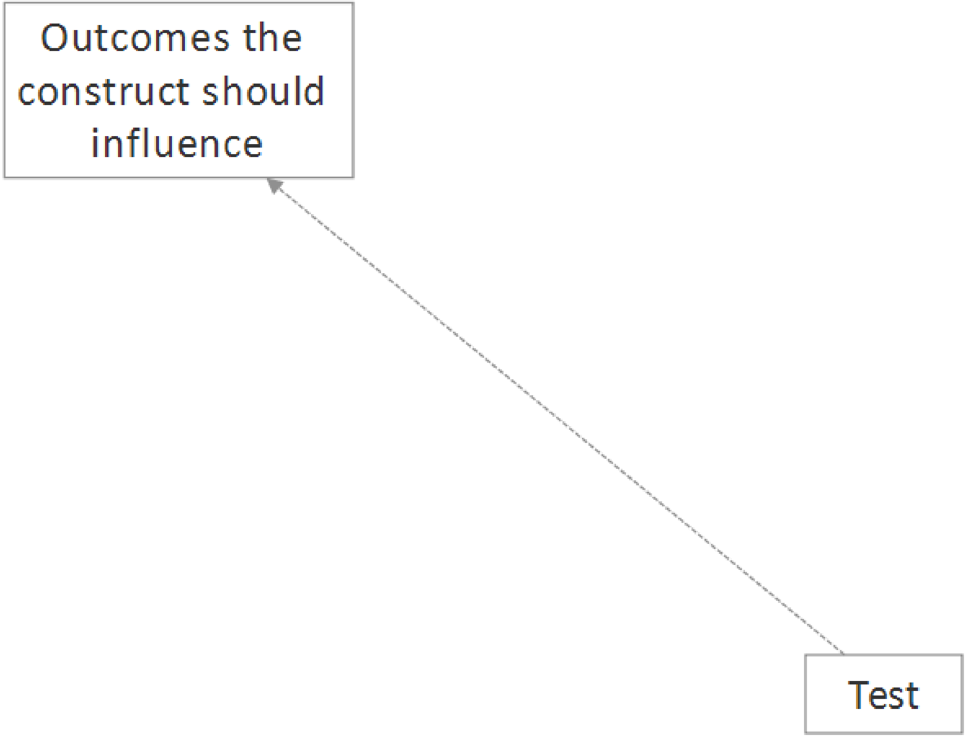
From Wikipedia, the free encyclopedia

For other uses, see Reliability.

Reliability in statistics and psychometrics is the overall consistency of a measure.[1] A measure is said to have a high reliability if it produces similar results under consistent conditions. "It is the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores. Scores that are highly reliable are accurate, reproducible, and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test takers, essentially the same results would be obtained. Various kinds of reliability coefficients, with values ranging between 0.00 (much error) and 1.00 (no error), are usually used to indicate the amount of error in the scores." [2] For example, measurements of people's height and weight are often extremely reliable.[3][4]

https://en.wikipedia.org/wiki/Reliability_(statistics)


Validity
Outcomes the construct should influence
Predictive validity
Construct the test is supposed to measure
Reference standard
Construct validity
Criterion validity
Latent variable underlying test
Internal consistency reliability
Test
Reliability (e.g., inter-rater, test-retest)
Reliability

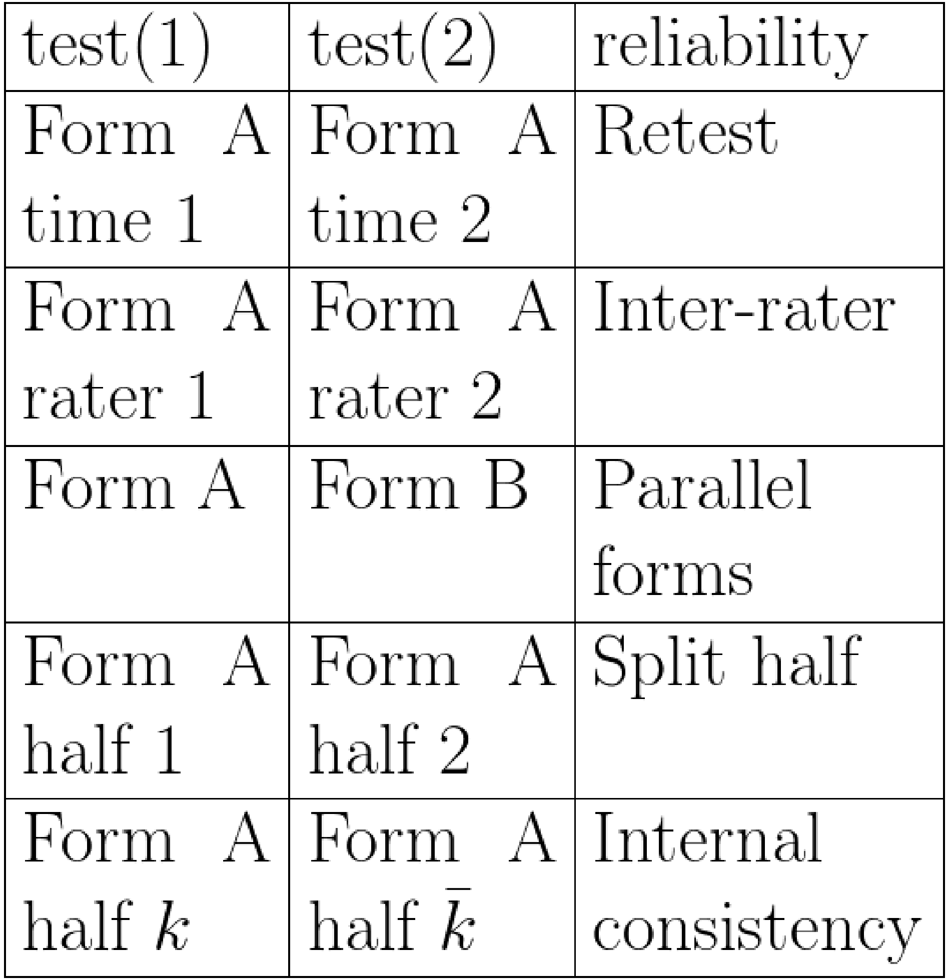


Outcomes the construct should influence

Test

Utility

Basics



test(1)	test(2)	reliability
Form A time 1	Form A time 2	Retest
Form A rater 1	Form A rater 2	Inter-rater
Form A	Form B	Parallel forms
Form A half 1	Form A half 2	Split half
Form A half k	Form A half $\bar{k}$	Internal consistency

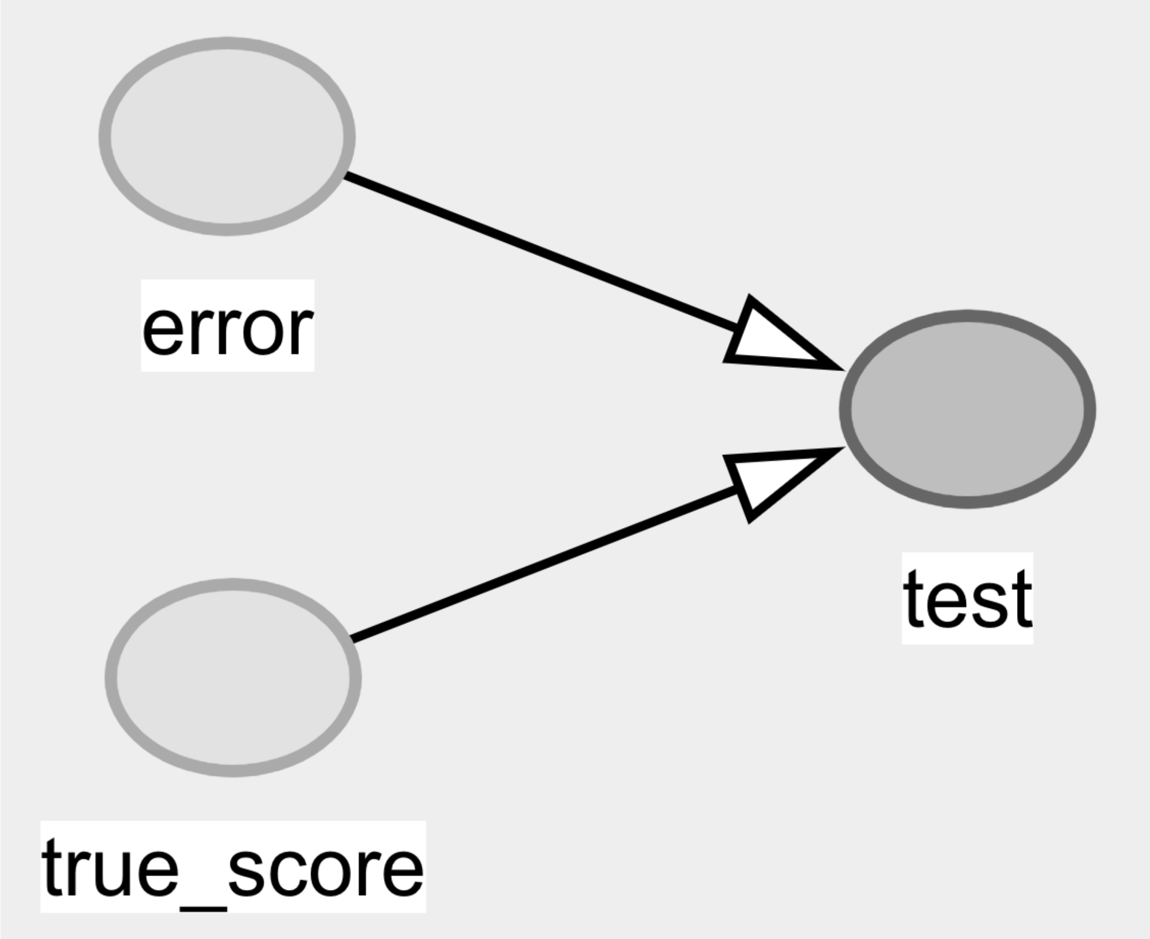
Classical test theory notion of reliability

Observed *test score* is a function of a true score and (random) *error*

The *true score* is the score that would be obtained on an arbitrarily large number of repeated assessments under identical conditions

A test score reliability is the fraction of the variance in the test score that is attributable to the true score

Since we don’t observe test scores under an arbitrary large number of repeated observations, we have to do some tricks to get an estimate of reliability

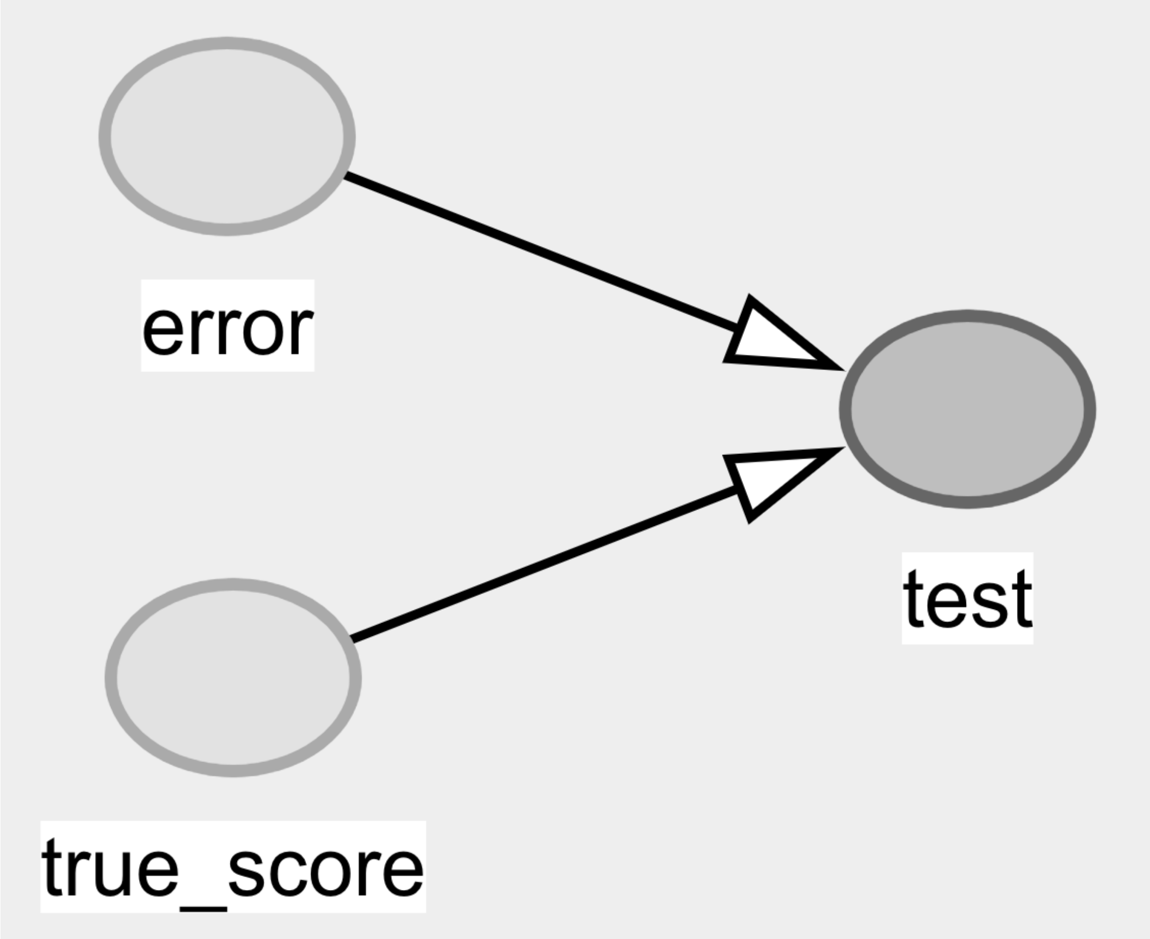


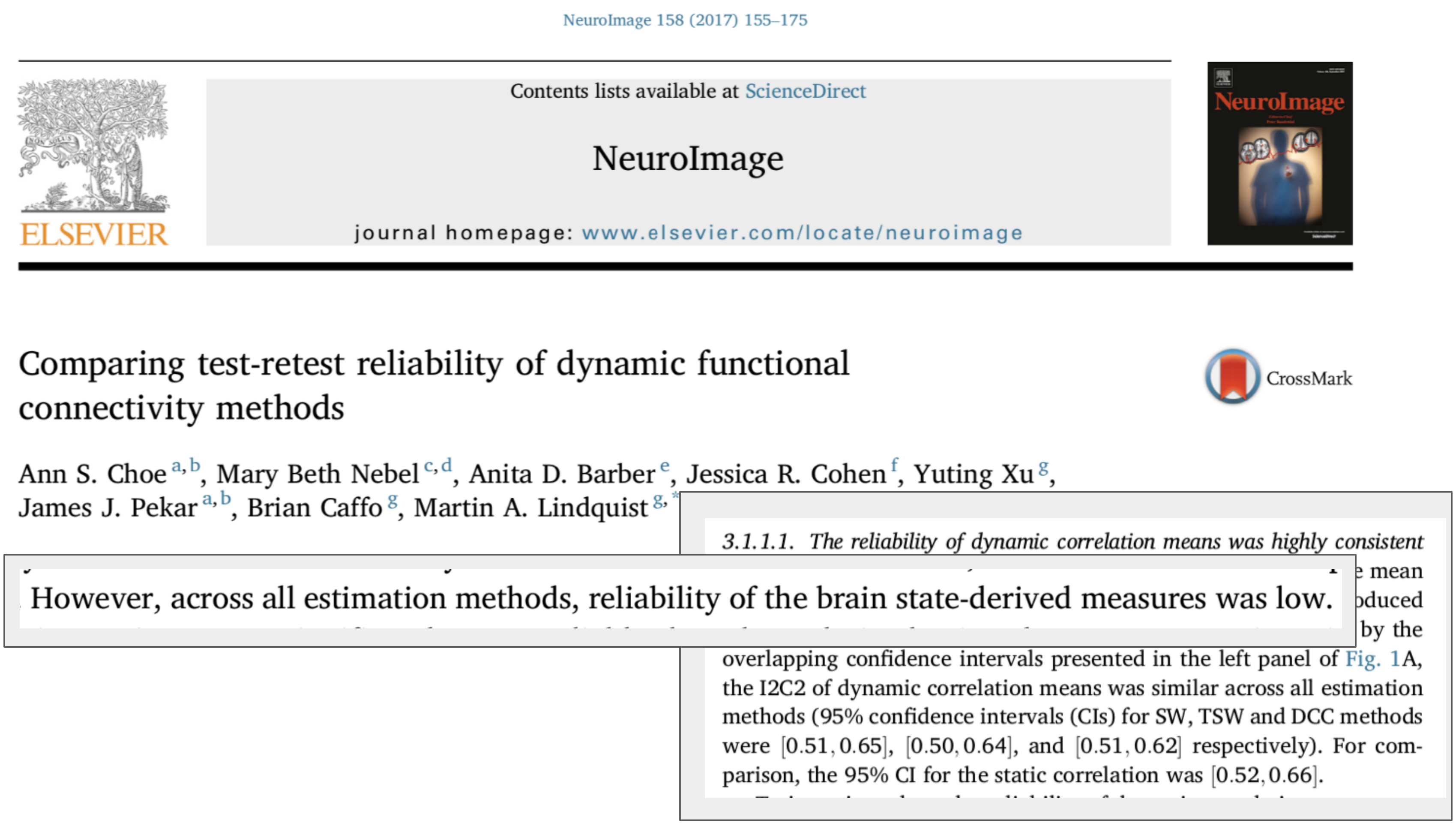
Classical test theory notion of reliability

$$x = t + e$$

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

$$\rho_{xx'} = \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$





NeuroImage 158 (2017) 155–175

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Comparing test-retest reliability of dynamic functional connectivity methods

Ann S. Choe[a,b], Mary Beth Nebel[c,d], Anita D. Barber[e], Jessica R. Cohen[f], Yuting Xu[g], James J. Pekar[a,b], Brian Caffo[g], Martin A. Lindquist[g,*]

However, across all estimation methods, reliability of the brain state-derived measures was low.

3.1.1.1. The reliability of dynamic correlation means was highly consistent

e mean

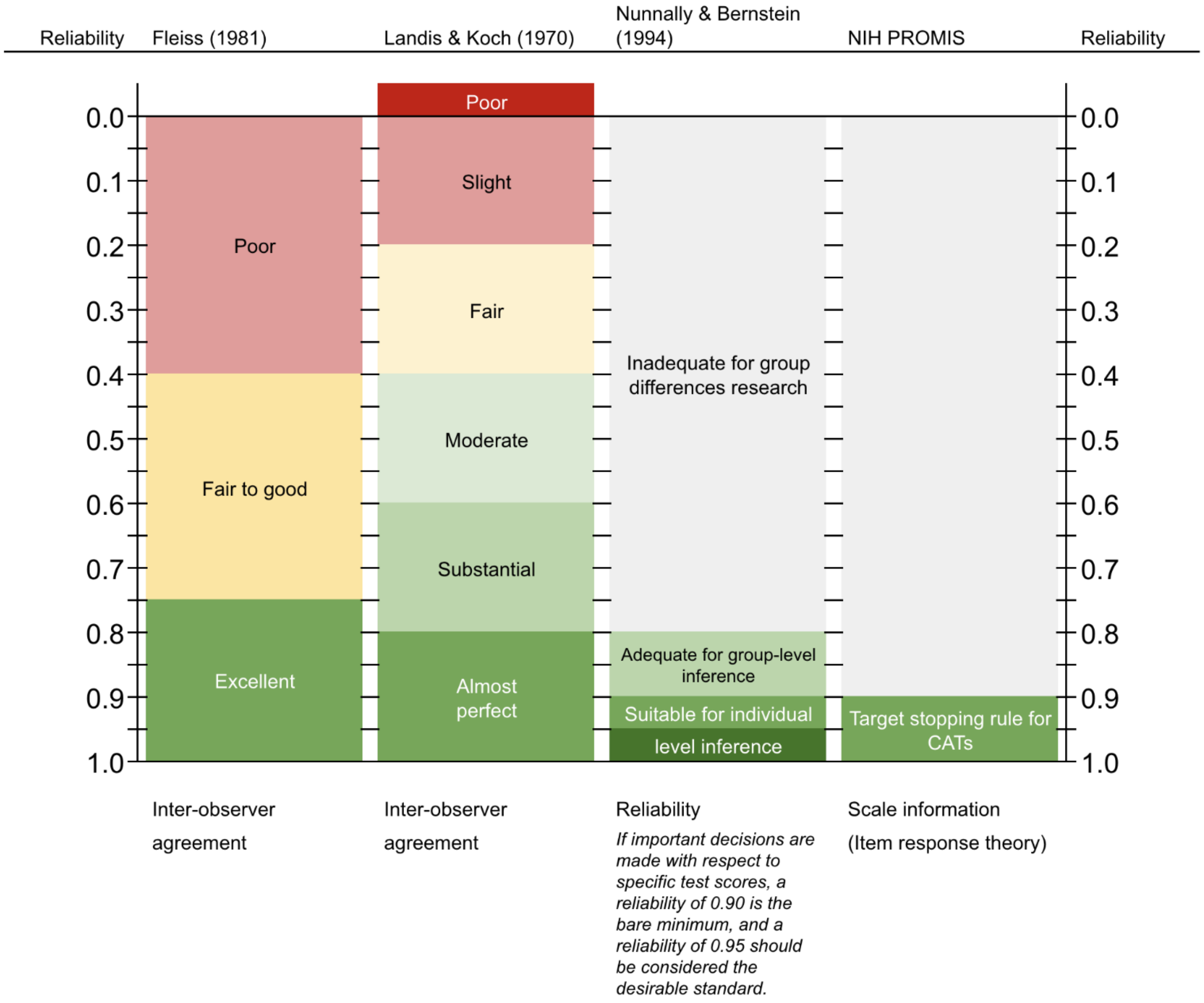
oduced

by the

overlapping confidence intervals presented in the left panel of Fig. 1A, the I2C2 of dynamic correlation means was similar across all estimation methods (95% confidence intervals (CIs) for SW, TSW and DCC methods were $[0.51, 0.65]$, $[0.50, 0.64]$, and $[0.51, 0.62]$ respectively). For comparison, the 95% CI for the static correlation was $[0.52, 0.66]$.

Implications

Reliability standards



Inter-observer agreement

Inter-observer agreement

Reliability

If important decisions are made with respect to specific test scores, a reliability of 0.90 is the bare minimum, and a reliability of 0.95 should be considered the desirable standard.

Scale information (Item response theory)

Study design: sample size

Sample size

Lehr’s equation - number needed per group (n) to detect a standardized effect size (d) with type-I error level of 5% and type-II error level of 20% (16)

$$n = \frac{16}{d^2}$$

If d = .5 , n = 64

Imagine you are planning a study

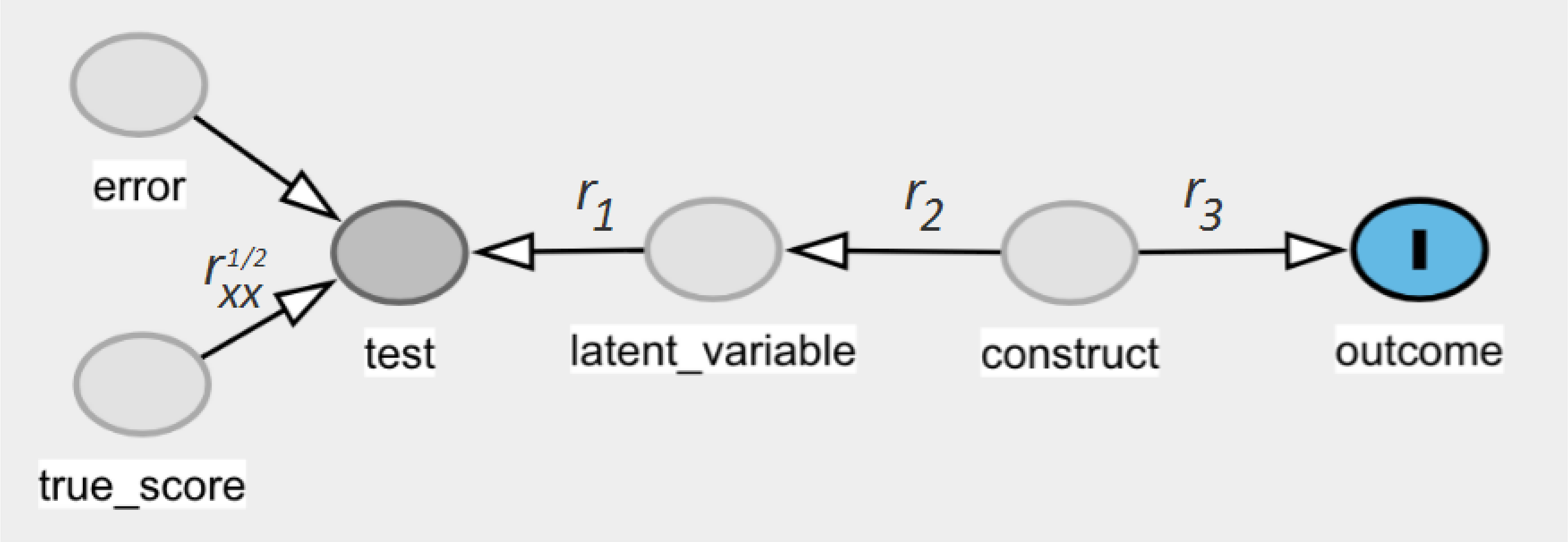
You have a treatment that can produce a 0.5 SD difference in the means *of the true score* across treated and non-treated participants (d = 0.5)

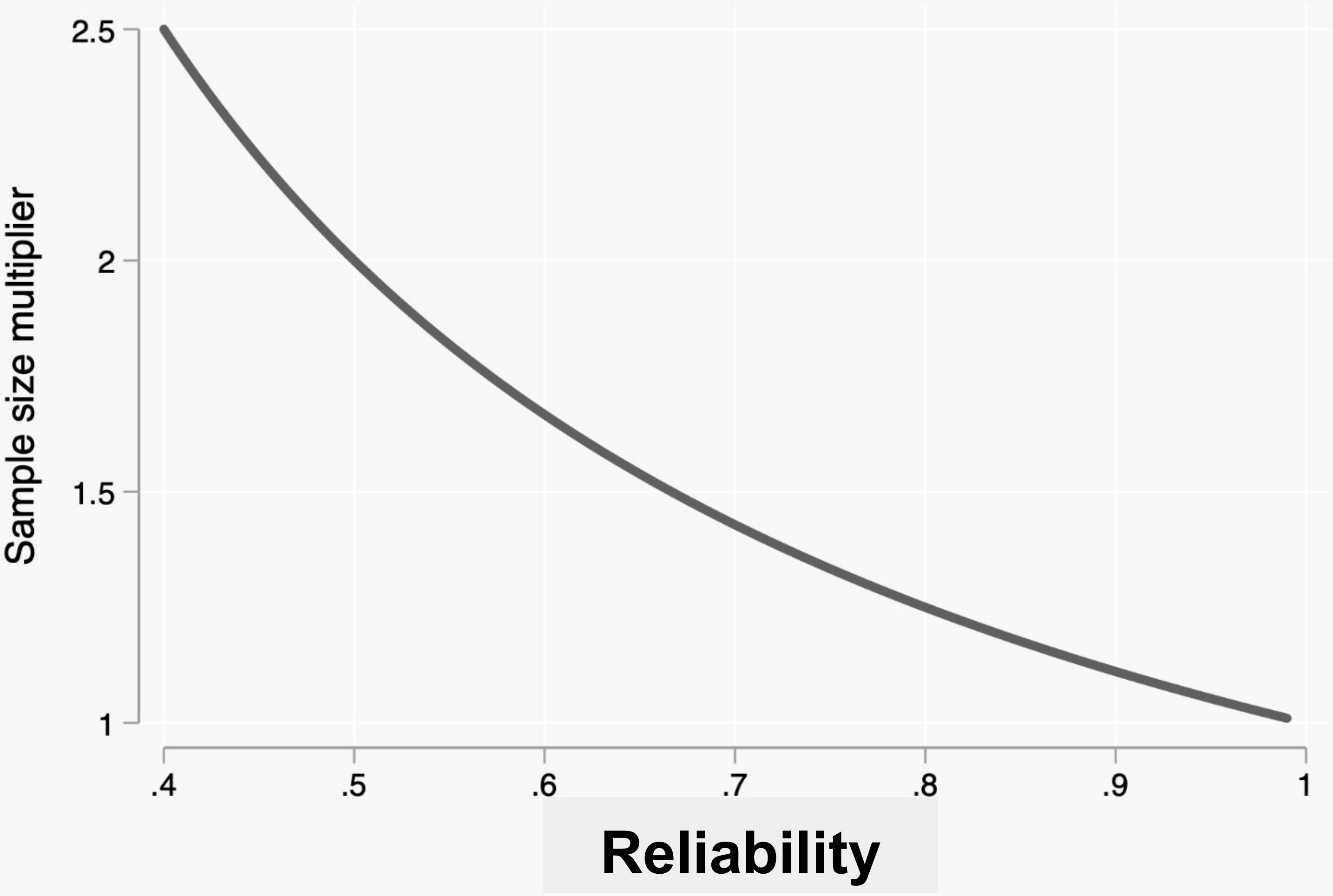
You have a measure of the outcome that is perfectly reliable (*REL* = 1.0)

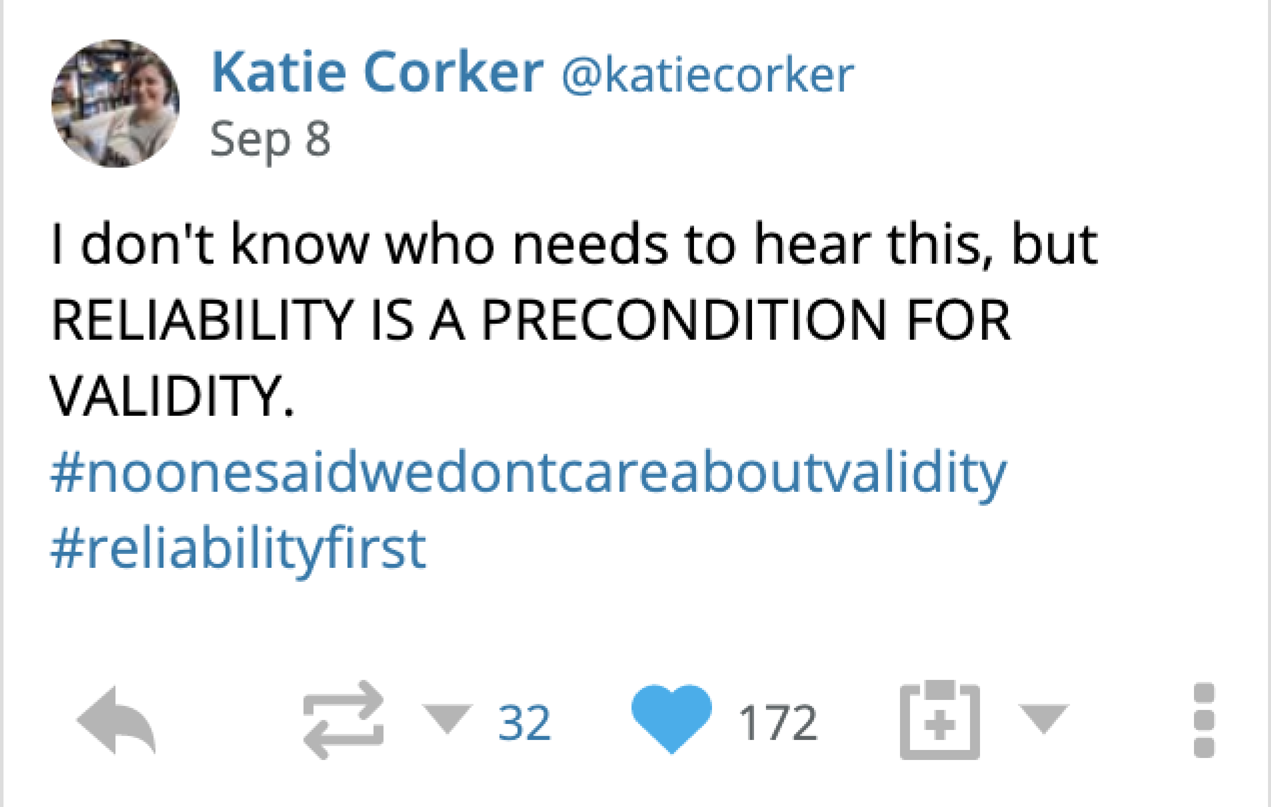
How many people do you need to randomize per group to have 80% power with a two sided type-I error level of 5%

$$n = \frac{16}{d^2} \times \frac{1}{REL}$$

If
 d = .5 , r = .7
then
 n = 92


error
$r_{xx}^{1/2}$
true_score
test
r_1
latent_variable
r_2
construct
r_3
outcome


2.5
2
1.5
1
Sample size multiplier
.4
.5
.6
.7
.8
.9
1
Reliability



Katie Corker @katiecorker
Sep 8

I don't know who needs to hear this, but RELIABILITY IS A PRECONDITION FOR VALIDITY. #noonesaidwedontcareaboutvalidity #reliabilityfirst

Clinical work

Imagine you are a clinician: a geriatrician (1)

1. You want to identify older adults who would be unsafe drivers because of cognitive impairment, using a test of mental status
2. You decide that adults in the worst quintile of *true cognitive status* would be unsafe drivers
3. You want to make the right decision 9 times out of 10

How reliable does your test have to be?

Imagine you are a clinician: a geriatrician (2)

Answer:

If you consider “true positive” and “true negatives” as correct decisions (i.e., you would like a “hit rate” of 90%), you need a reliability of .8

The PPV when reliability = .8 is about 75%

Imagine you are a clinician: a geriatrician (3)

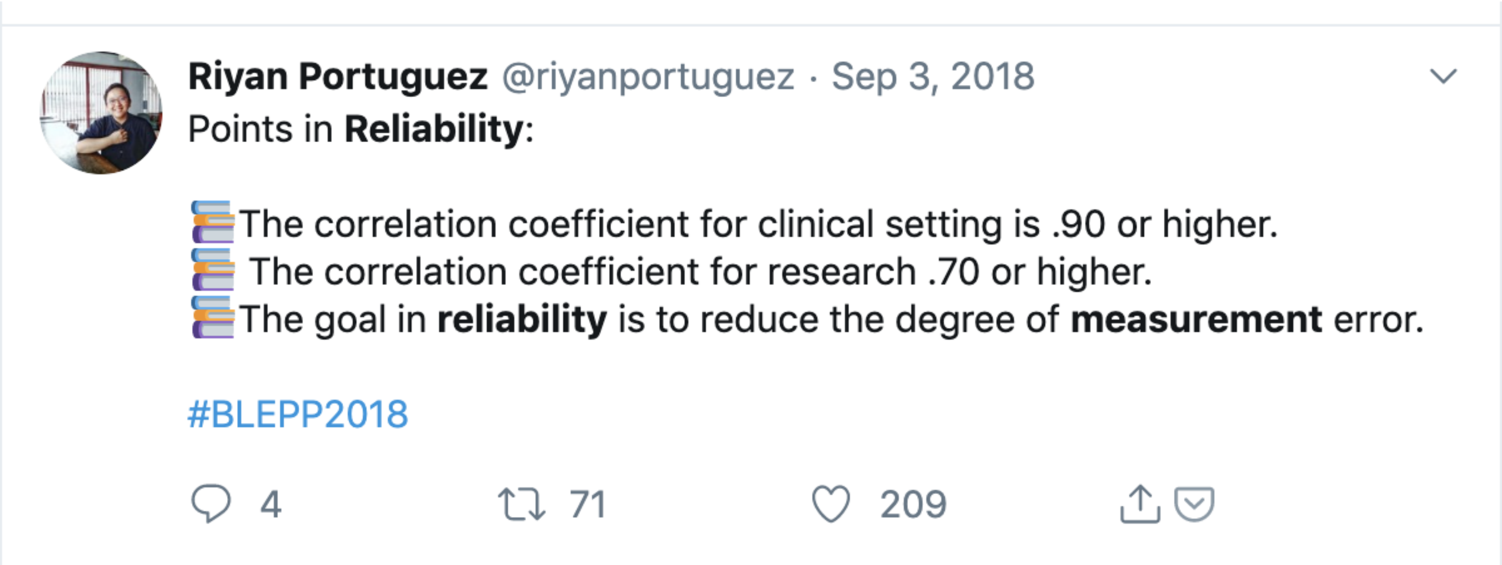
Answer:

If you only consider “true positive” as correct decisions (i.e., you would like a positive predictive value of 90%), you need a reliability of .97 (!)

The PPV when reliability = .8 is about 75%

- Standards for reliability depend on lots of things
- In research, highly reliable measures can improve power
- In clinical settings, highly reliable measures are the foundation of good clinical practice

MAJOR MESSAGE



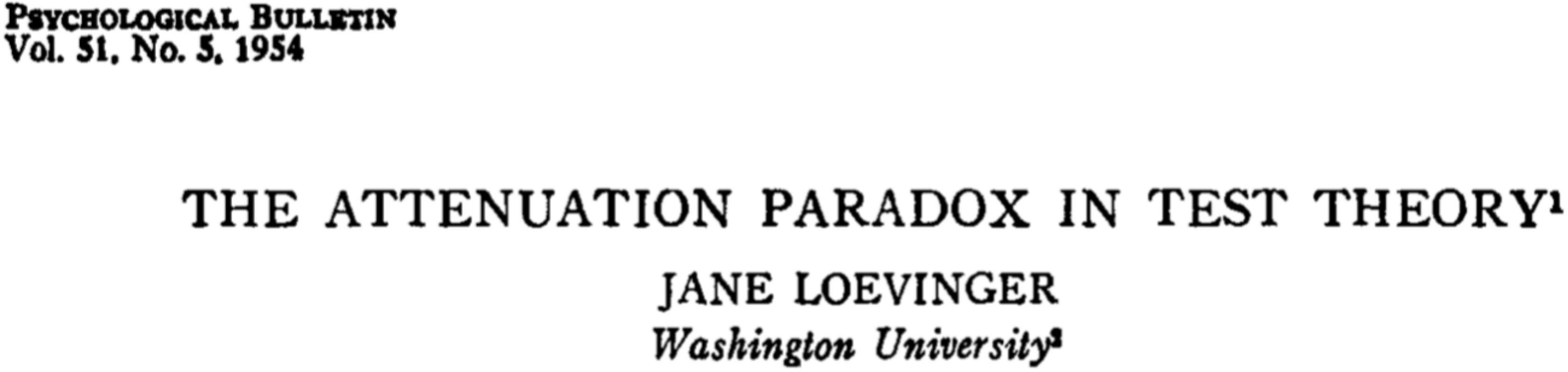
Riyan Portuguez @riyanportuguez · Sep 3, 2018

Points in **Reliability**:

📚The correlation coefficient for clinical setting is .90 or higher.
📚 The correlation coefficient for research .70 or higher.
📚The goal in **reliability** is to reduce the degree of **measurement** error.

#BLEPP2018

Paradox



PSYCHOLOGICAL BULLETIN
Vol. 51, No. 5, 1954

THE ATTENUATION PARADOX IN TEST THEORY[1]

JANE LOEVINGER
Washington University[2]

PSYCHOLOGICAL BULLETIN
Vol. 51, No. 5, 1954

THE ATTENUATIO

J

W

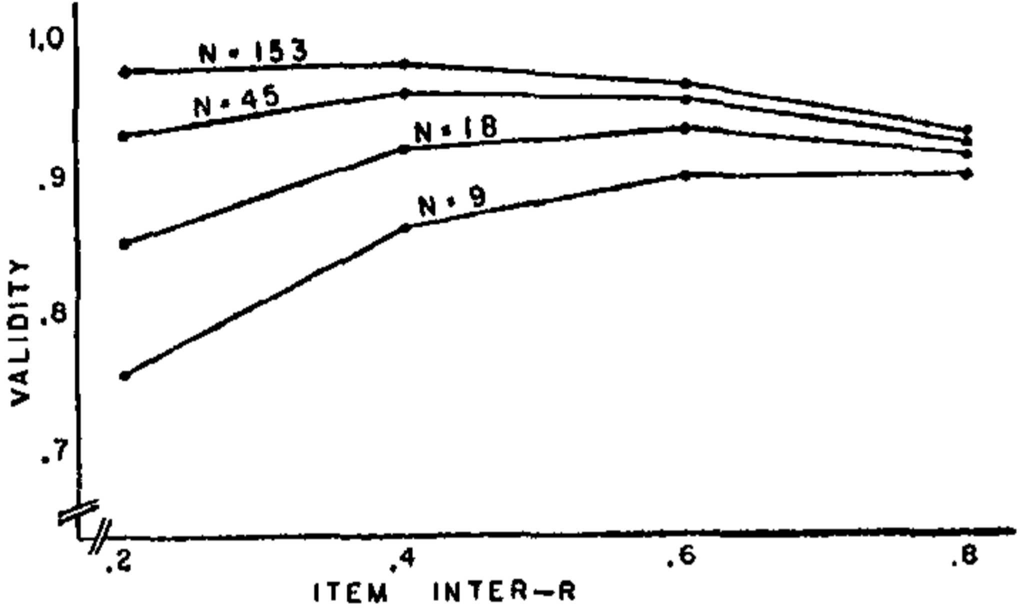


Fig. 1. A ttenuation Paradox as a Function of Number of Items for Tests Composed of Median Equivalent Items. Data from Brogden's (1) Table 2.

The real message behind the Attenuation Paradox is: the CTT notion of reliability is limited, and, wrong. Should use item response theory (IRT) instead.

Optimizing

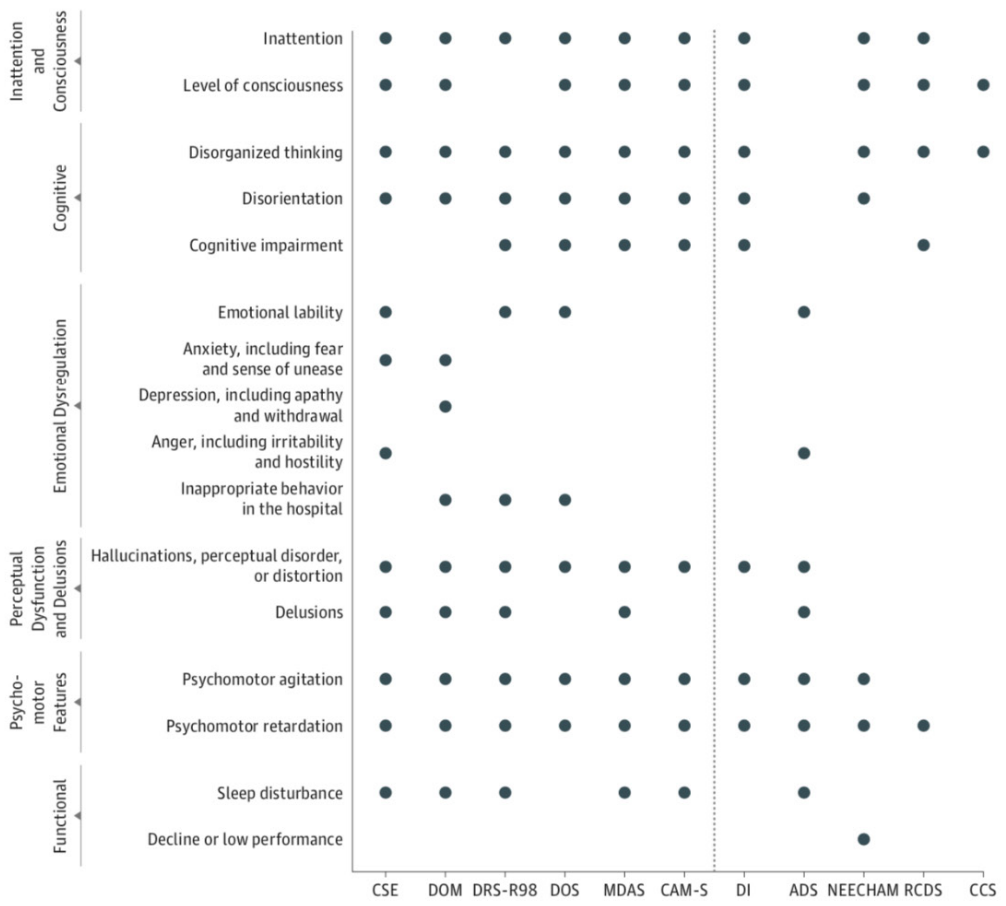
How to optimize reliability of measures

1. Use rigorously designed and validated measures (c.f., COSMIN)
    - Be critical and skeptical in your review of literature. Look out for
        - “Bloated specific” measures that sacrifice fidelity for bandwidth
        - Forms pruned with “alpha if item deleted”
    - Pilot test instruments in samples from your target population
2. Resist the urge to use *fixed* short forms
3. Satisfy the urge to use short forms with Computerized Adaptive Tests
4. Staff training and continual quality assessment in conducting research

Assessment of Instruments for Measurement of Delirium Severity: A Systematic Review

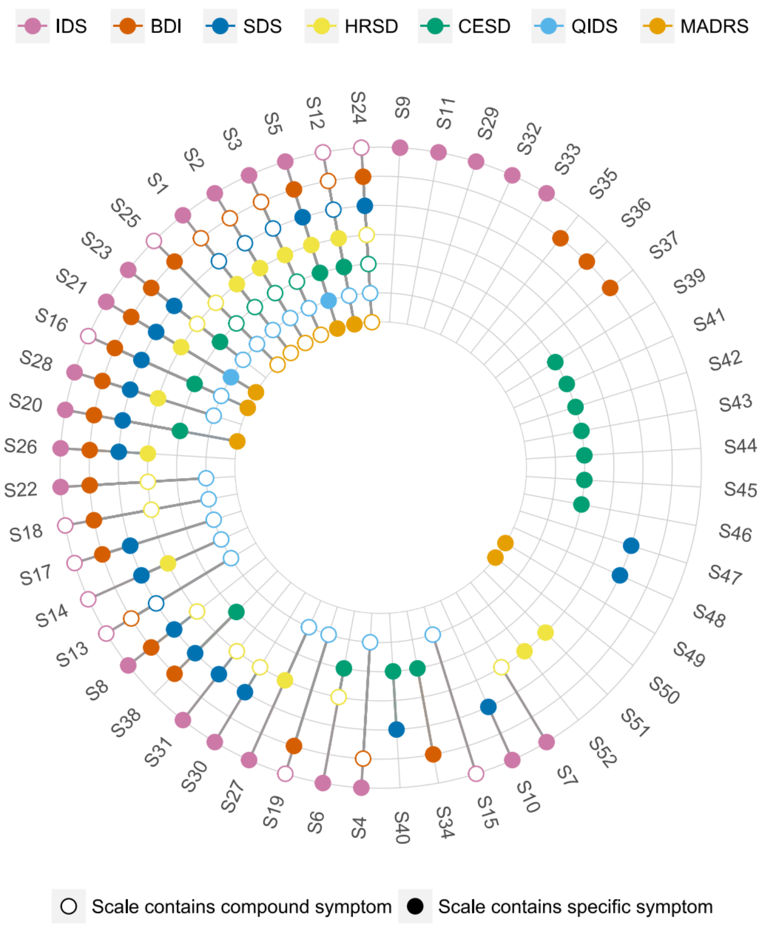
Jones RN et al. JAMA Intern Med. 2019:179(2);231-239

Figure 2. Domain Coverage of 11 Multi-item Delirium Severity Instruments



		CSE	DOM	DRS-R98	DOS	MDAS	CAM-S	DI	ADS	NEECHAM	RCDS	CCS
Inattention and Consciousness	Inattention	●	●	●	●	●	●	●		●	●	
	Level of consciousness	●	●		●	●	●	●		●	●	●
Cognitive	Disorganized thinking	●	●	●	●	●	●	●		●	●	●
	Disorientation	●	●	●	●	●	●	●		●		
	Cognitive impairment			●	●	●	●	●			●	
Emotional Dysregulation	Emotional lability	●		●	●				●			
	Anxiety, including fear and sense of unease	●	●									
	Depression, including apathy and withdrawal		●									
	Anger, including irritability and hostility	●							●			
	Inappropriate behavior in the hospital		●	●	●							
Perceptual Dysfunction and Delusions	Hallucinations, perceptual disorder, or distortion	●	●	●	●	●	●	●	●			
	Delusions	●	●	●		●			●			
Psycho-motor Features	Psychomotor agitation	●	●	●	●	●	●	●	●	●		
	Psychomotor retardation	●	●	●	●	●	●	●	●	●	●	
Functional	Sleep disturbance	●	●	●		●	●		●			
	Decline or low performance									●		

Black dot indicates representation of a domain in the instrument; either partial or full coverage of a domain met criteria for inclusion by the expert panel. ADS indicates Agitation Distress Scale; CAM-S, Confusion Assessment Method–Severity Score; CCS, Communication Capacity Scale; CSE, Confusion State Examination; DI, Delirium Index; DOM, Delirium-O-Meter; DOS, Delirium Observation Screening; DRS-R98, Delirium Rating Scale-Revised-98; MDAS, Memorial Delirium Assessment Scale; NEECHAM, Neelon and Champagne Confusion Scale; RCDS, Recoverable Cognitive Dysfunction Scale.



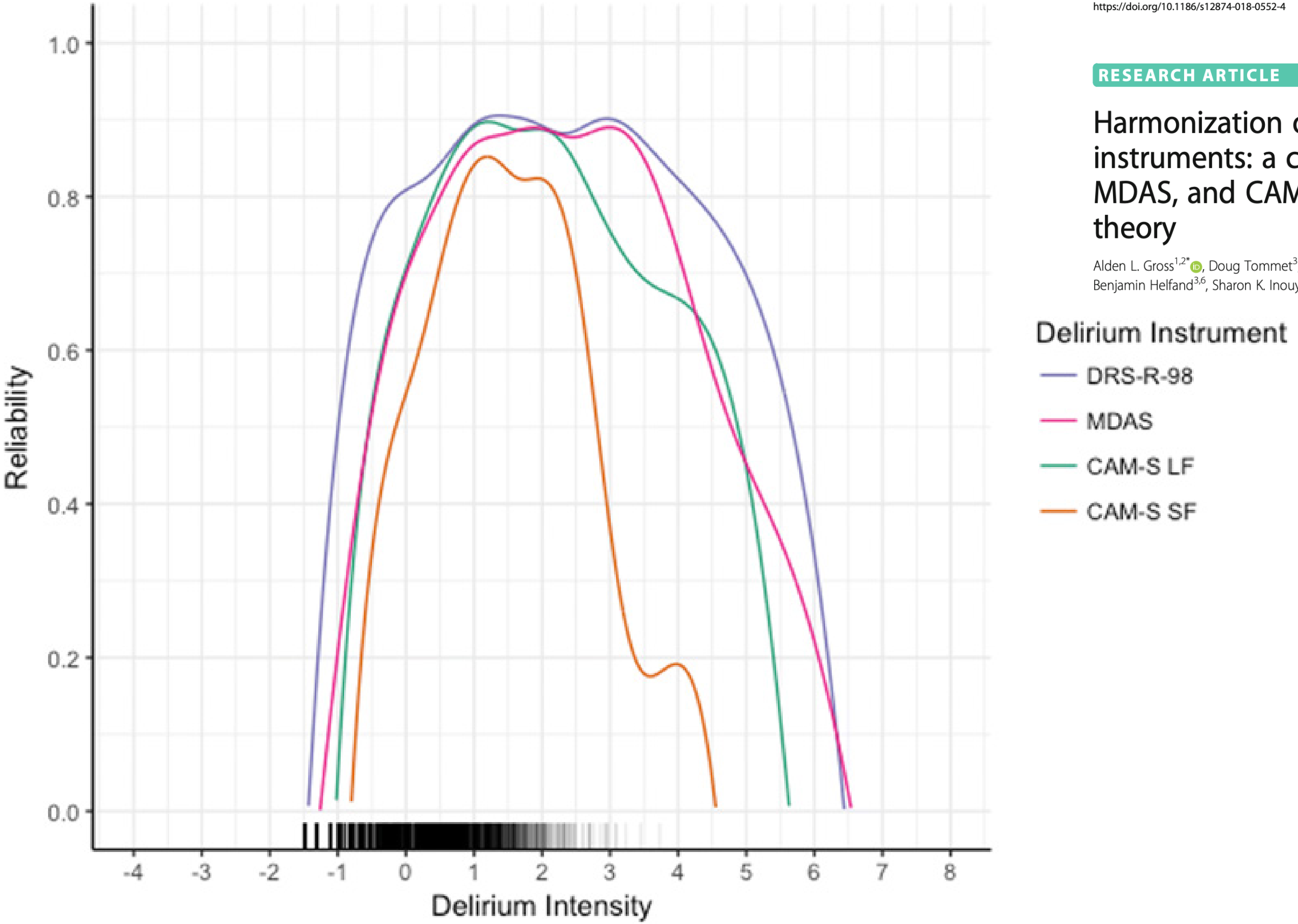
1 Early insomnia
2 Middle insomnia
3 Late insomnia
4 Hypersomnia
5 Sad mood
6 Anxious
7 Panic
8 Irritable
9 Mood reactivity
10 Diurnal variation
11 Grief
12 Appetite decrease
13 Appetite increase
14 Weight decrease
15 Weight increase
16 Concentration
17 Indecisiveness
18 Guilt
19 Worthlessness
20 Pessimism
21 Suicidal ideation
22 Interest loss
23 Pleasure loss
24 Fatigue
25 Energy loss
26 Libido
27 Retardation
28 Agitation
29 Somatic complaints
30 Sympathetic arousal
31 Gastrointestinal
32 Interpersonal sensitivity
33 Leaden paralysis
34 Past failure
35 Punishment
36 Self-dislike
37 Self-criticalness
38 Crying
39 Lonely
40 Effort
41 Talked less
42 People are unfriendly
43 People disliked me
44 Feeling bothered
45 Feeling good
46 Feeling happy
47 Feeling needed
48 Life is full
49 Inner tension
50 Inability to feel
51 Hypochondriasis
52 Loss of insight

The 52 symptoms of major depression: Lack of content overlap among seven common depression scales.

Fried EI. J Affect Disord. 2017.

https://www.ncbi.nlm.nih.gov/m/pubmed/27792962/

Fig. 1. Co-occurrence of 52 depression symptoms across 7 depression rating scales. Colored circles for a symptom indicate that a scale directly assesses that symptom, while empty circles indicate that a scale only measures a symptom indirectly. For instance, the IDS assesses item 4 hypersomnia directly; the BDI measures item 4 indirectly via a general question on sleep problems; and the SDS does not capture item 4 at all. Note that the 9 QIDS items analyzed correspond exactly to the DSM-5 criterion symptoms for MDD. Please see the online version for colors; in the black and white version, the circles respresent (from outer to inner circle): IDS, BDI, SDS, HRSD, CESD, QIDS, and MADRS.



Gross *et al. BMC Medical Research Methodology* (2018) 18:92
https://doi.org/10.1186/s12874-018-0552-4

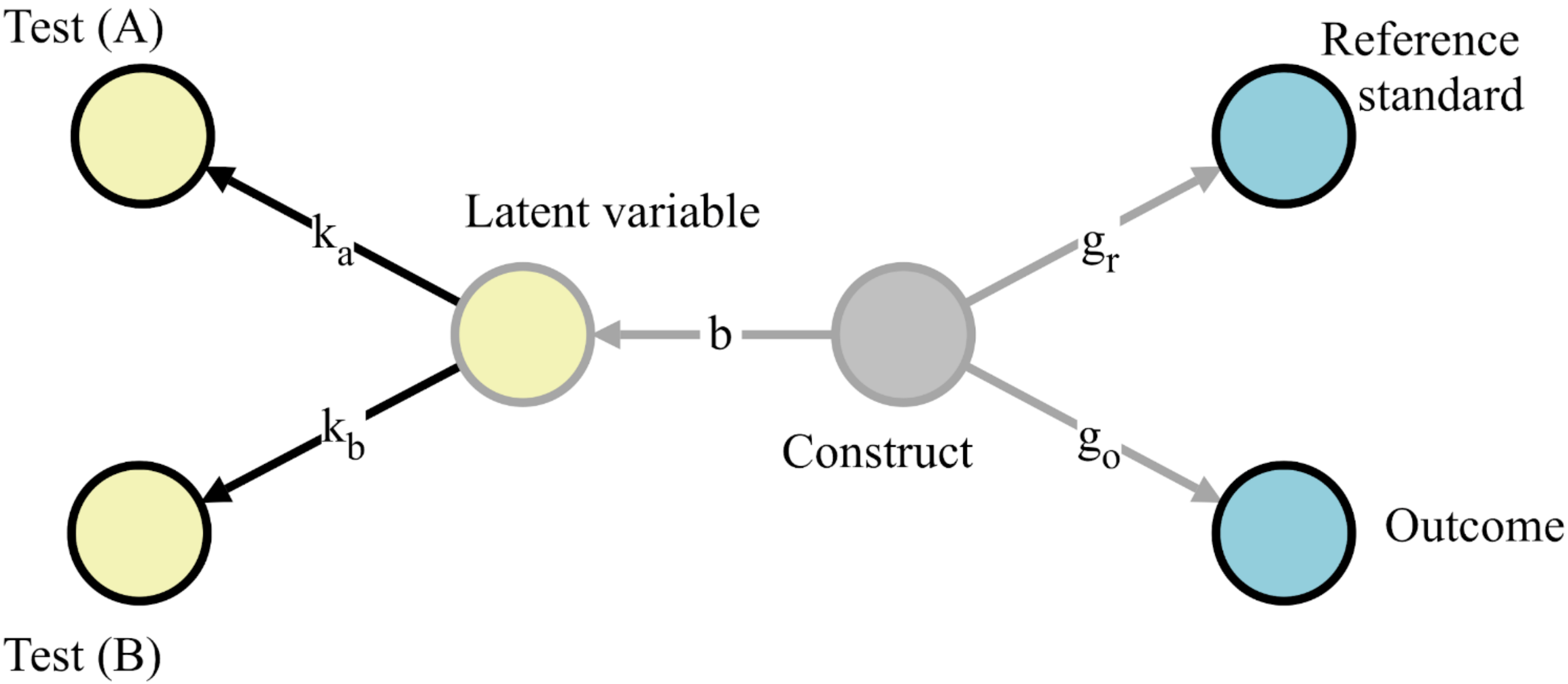
BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Harmonization of delirium severity instruments: a comparison of the DRS-R-98, MDAS, and CAM-S using item response theory

Alden L. Gross[1,2*], Doug Tommet[3], Madeline D'Aquila[4], Eva Schmitt[4], Edward R. Marcantonio[4,5], Benjamin Helfand[3,6], Sharon K. Inouye[4,5†], Richard N. Jones[3†] and for the BASIL Study Group

Discussion and questions


Test (A)
Reference standard
Latent variable
k_a
g_r
b
k_b
Construct
g_o
Outcome
Test (B)