A gentle introduction to reproducibility and validity

Henrik Hein Lauridsen
Associate Professor
Research unit for Clinical Biomechanics

SDU

1

You will see…

"More on this in the advanced course"

This course is ONLY an overview.

If you need to know it in more details, you need the advanced course.

SDU

2

COSMIN definitions

Overview of COSMIN definitions

Download it from www.clinimetrics.sdu.dk

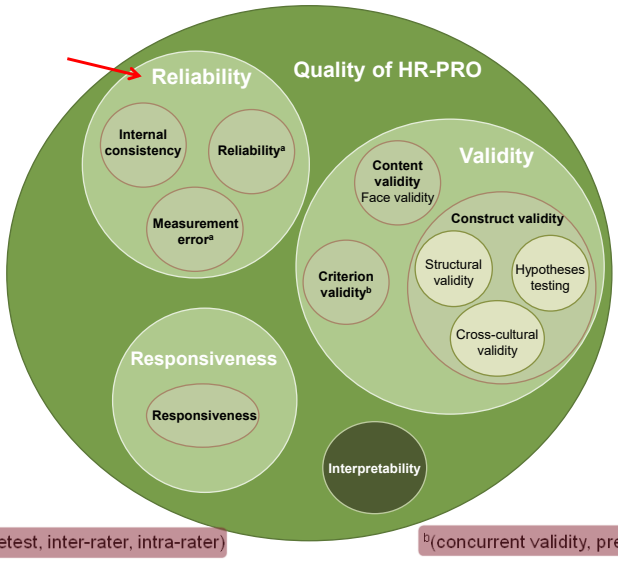
COSMIN definitions of domains, measurement properties, and aspects of measurement properties

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
	Criterion validity		The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores

† *The word 'true' must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy (ref Streiner & Norman)*
* *Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument*

3

The COSMIN taxonomy

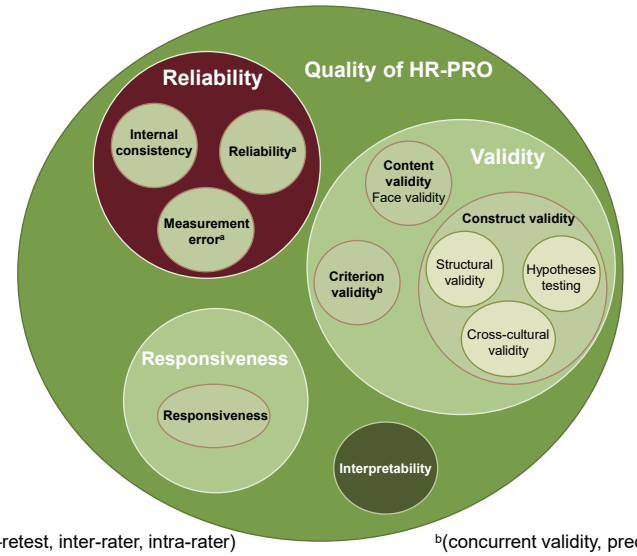


[a](test–retest, inter-rater, intra-rater)

[b](concurrent validity, predictive validity)

4

The COSMIN taxonomy



[a](test–retest, inter-rater, intra-rater)

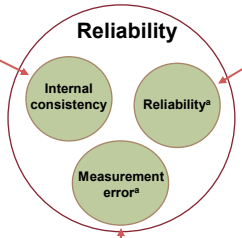
[b](concurrent validity, predictive validity)

Definitions

The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions

The degree of the interrelatedness among the items

Items are repeated measurements but cross-sectional data

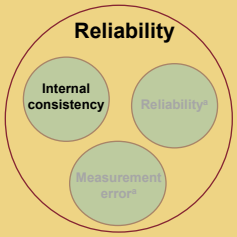


The proportion of the total variance in the measurements which is due to 'true' differences between patients

Scores on repeated measurements (e.g. test-retest design)

The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured

Internal consistency



SDU

7

Internal consistency

"Degree of inter-relatedness among items"

Used on multi-item questionnaires

Assumes that all items measure the same concept

Measured at one point in time → Is a reliability measure as the repetition of measurements are the different items in the scale

De Vet et al. Measurement in Medicine. A Practical Guide. 2011

SDU

8

Why do we need an internally consistent scale?

It makes interpretation of a composite score (sum score) as a reflection of the items in the test possible

Implications

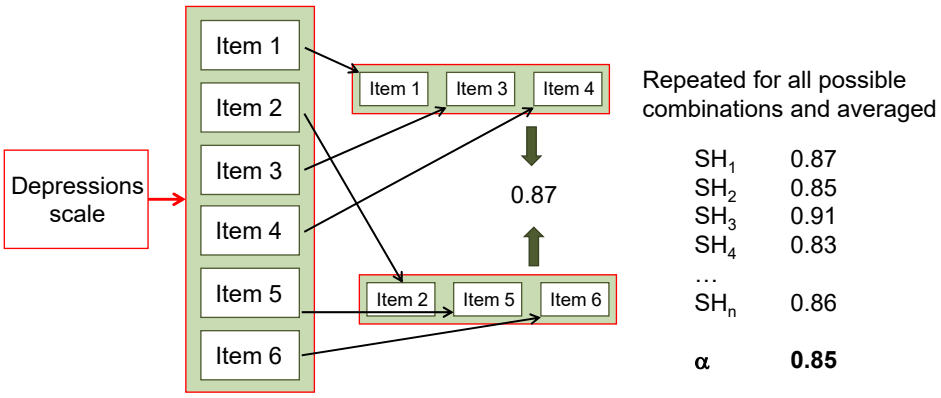
1. The items should be moderately correlated with each other
2. Each should correlate with the total scale score

Also used for item reduction

- Advanced course…

Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. Fifth edition. Oxford Medical Publications, 2015.

Cronbach's α



Repeated for all possible combinations and averaged

SH_1	0.87
SH_2	0.85
SH_3	0.91
SH_4	0.83
…	
SH_n	0.86
α	**0.85**

Cronbach's α

A measure of internal consistency (polytomous items only)

Interpretation

α = 0 = no items are related i.e. no correlation
α = 1 = all items are identical i.e. perfect correlation
α = 0.7-0.9 = often accepted as satisfactory for most scales (COSMIN: α > 0.7)

NB: α is often high in multi-dimensional scales, therefore establish structural validity first, then internalt consistency*

* Cortina J. What is coefficient alpha: an examination of theory and applications. Journal of applied psychology. 1993;78:98-104.

11

Cronbach's α at 'item level'

Item	Alpha
Item 1	0.73
Item 2	0.75
Item 3	0.69
…	
Item 15	0.86
Item 16	0.76
Test scale	**0.74**

Alpha excluding the one item

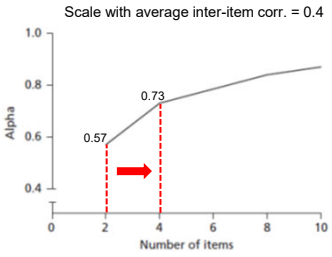
If substantially *higher* than Chronbach's α → maybe misfitting

12

Cronbach's α

Cronbach's α is dependent on the sample size and number of items:

- ↑ sample size → ↑ Cronbach's α
- ↑ number of items → ↑ Cronbach's α



Item per subscale	Rating	Sample size N < 100	N = 100-300	N > 300
≤ 6	Excellent	.75	.80	.85
	Good	.70	.75	.80
	Moderate	.65	.70	.75
	Fair	.60	.65	.70
7-11	Excellent	.80	.85	.90
	Good	.75	.80	.85
	Moderate	.70	.75	.80
	Fair	.65	.70	.75
≥ 12	Excellent	.85	.90	.90
	Good	.80	.85	-
	Moderate	.75	.80	.85
	Fair	.70	.75	.80

An internal consistency coefficient falling below the "Fair" rating for its particular cell would be deemed "Unsatisfactory"

Streiner & Norman. Health Measurement Scales. A practical guide to their development and use. Fifth edition.
Ponterotto, J.G. and Ruckdeschel, D. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. Perceptual and Motor Skills, 105, 997–1014.

13

Assumptions of Cronbach's α

- Unidimensionality (homogeneity of items) → Structural validity established BEFORE internal consistency
- Item scorer are correlated (i.e. no items are irrelevant)
- Responses to items are normally distributed
- Items have equal variances
- Items have equal factor loading
  - Called 'Essential tau-equivalence'

Seldom met

Result:
- Scales with true high internal consistency may be rejected
- Scales with true low internal consistency may be retained

Stensen K, Lydersen S. Internal consistency: from alpha to omega? *Tidsskrift for Den norske legeforening* Published Online First: 29 August 2022

14

A better coefficient – McDonald's Ω

- Based on factor analytical techniques
- Does not assume Tau-equivalence
- Differences between α and Ω are often small, **BUT** can be large
- Used less frequently as more complex to calculate statistically
- Ω should (probably) be presented alongside α

15

Reliability & measurement error

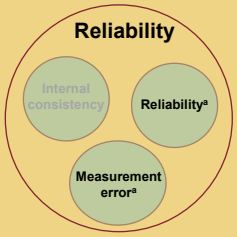
Two related but distinct measurement properties

Based on the same data:

Repeated measurements in stable patients

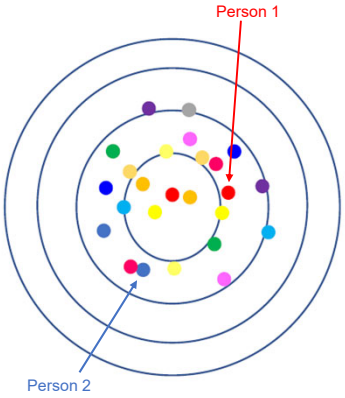
Refers to different concepts

Assessed with different statistical formulas



16

Measurement error

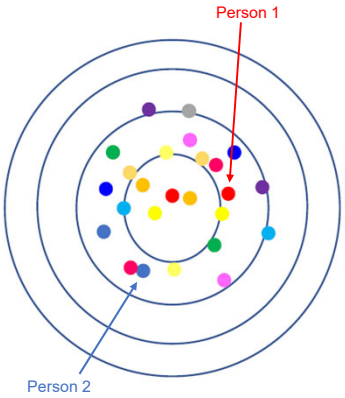


Precision of the score

How close are the scores of repeated measurements in a stable patient to each other?

17

Measurement error



Precision of a score when using it in clinical research or practice

- It refers to variation in scores within a person
- Assumption: consistent across the scoring continuum
- Expressed in unit of measurement

18

Reliability

Is the ability of an instrument to distinguish between people (despite measurement error)

- True variation between patients
- Variation due to other sources (i.e. error)

Other sources

- Type of machine
- Variation between raters
- Variation in occasion
- Variation in software
- Etc.

A coefficient between 0 – 1 → hard to interpret

19

Reliability – other sources

Components	Sources of variation that can influence the score
Equipment & preparation	 Questionnaire forms, computers, stair steps of a specific height, device (e.g. stopwatch, probe, tube), machine, gels, MRI scanner, training, instructions, etc.
Collection of raw data	 All actions undertaken by patient and professional(s) to collect the data, before any data processing
Data processing & storage	 All actions undertaken on the raw data to store it in a usable (electronic) form for later data manipulation
Assignment of the score	 Methods used to convert processed data into a score that constitutes the outcome measurement instrument

Mokkink, Lidwine B, Iris Eekhout, Maarten Boers, Cees Pm Van Der Vleuten, and Henrica Cw De Vet. 'Studies on Reliability and Measurement Error of Measurements in Medicine – From Design to Statistics Explained for Medical Researchers'. *Patient Related Outcome Measures* Volume 14 (July 2023): 193–212.

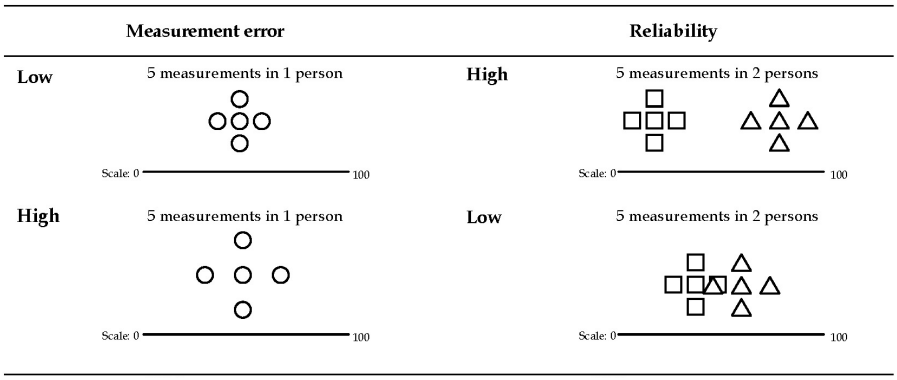
20

Reliability – other sources

- Which source influence the score the most?
- Can this source be better standardized?
- Should it be restriced to improve the measurement?

21

Measurement error & reliability



De Vet et al. Measurement in Medicine. A Practical Guide. 2011

22

Measurement error & reliability

Measurement level	Parameter
Measurement error	
Continuous	Standard Error of the Measurement (SEM) Bland & Altman - Limits of agreement (LOA)
Ordinal	Proportion of specific agreement
Nominal	Proportion of specific agreement
Reliability	
Continuous	ICC
Ordinal	ICC or weighted kappa
Nominal	Unweighted kappa

De Vet et al. Measurement in Medicine. A Practical Guide. 2011

23

Measurement error

Standard error of the measurement (SEM)

"Is the standard deviation of errors of measurement that is associated with 'test' scores"

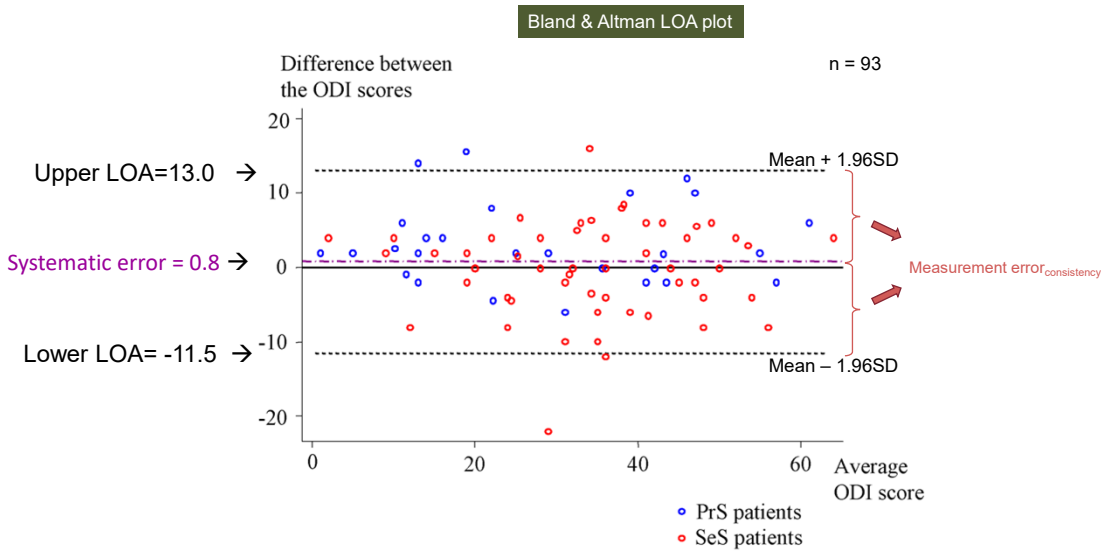
$$SEM = \sqrt{\sigma^2_{error}} = measurement\ error$$

More on this in the advanced course…

De Vet et al. 2006

24

Measurement error



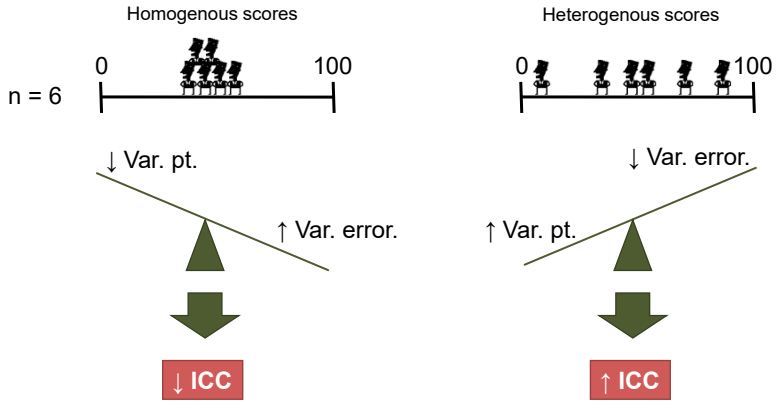
25

Reliability

Patient variance:
- Variability between persons i.e., variance in the patient scores (inter/intra-rater, time)

$$ICC = \frac{Var_{patient}}{Var_{patient} + Var_{error}}$$

Error variance:
- Systematic differences between e.g. observers (inter/intra-rater, time)
- Random error



Shrout & Fleis, 1979; McGraw & Wong, 1996

26

Reliability

$$ICC_{\text{consistency}} = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma_{error}}$$

Random error

$$ICC_{agreement} = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + (\sigma^2_{observers} + \sigma^2_{residual})}$$

Systematic error

Can be:
- Over time (test-retest)
- Different persons on same occasion (inter-rater)
- Same person on different occasions (intra-rater)

Random error

Test-retest
Inter-rater
Intra-rater

More on this in the advanced course…

Shrout & Fleis, 1979; McGraw & Wong, 1996

27

Reliability (reproducibility)

If this is how you feel right now…

The exercise will help you…

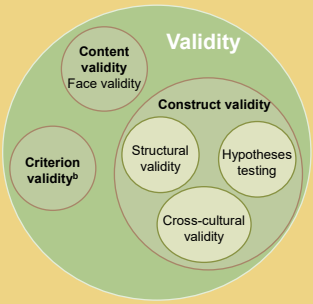
AND

the advanced course will too.

28

Validity

- **Face & content validity**
  - General considerations
  - What is the Content Validity Index?
  - Content validity and context
- **Criterion validity**
- **Construct validity**
  - Structural validity
  - Hypotheses testing



SDU

29

Face validity

'THE DEGREE TO WHICH A MEASUREMENT INSTRUMENT, INDEED, LOOKS AS THOUGH IT IS AN ADEQUATE REFLECTION OF THE CONSTRUCT TO BE MEASURED'

- **First impression: Does the scale appear to measure what it clamis to measure?**
- **Important:**
  - If no face validity → strong argument for NOT using the instrument

SDU

30

Content validity

'The degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured'

- Are the items a 'true' reflection of the relevant aspects of the construct?
- Are all items relevant for the target population?
- Are all items understandable for the target population?

COSMIN 2011

31

Content validity considerations

1. *Consider information about construct and situation*
   - *Specify construct to be measured clearly*
   - *What is the target population and purpose*
   - *Where in a conceptual model do the domain items belong*

Terwee et al. (2018) COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study

32

Content validity considerations

2. *Consider information about content of the measurement instrument*

Item relevance

I. *Are the items relevant for the*
   a) construct of interest?
   b) target population?
   c) context of use?

I. *Are the response options appropriate?*

II. *Is the recall period appropriate?*

Comprehensiveness (i.e. er alt med?)

I. *Are all key concepts included?*

Comprehensibility (i.e. forstås alt?)

I. *Are the instructions understood?*

II. *Are the response options understood?*

III. *Are the item wording appropriate (”wordsmithing”)*

IV. *Do the respons options match the question*

Terwee et al. (2018) COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study

33

Content validity - howto

Search the literature and reuse what is already there

- E.g. PROMIS → http://www.healthmeasures.net/explore-measurement-systems/promis

Use an expert panel to assess the content

- For a PROM this could be experts in the field and/or patients/respondents

Use a strategy to link the content to the construct using a conceptual model

- E.g. WHO conceptual model (Cieza & Stucki (2002, 2005))
- Wilson & Cleary conceptual model

Terwee et al. (2018) COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study

34

The Content Validity Index (CVI)

Is an index quantifying content validity for multi-item scales

Uses 'experts' to evaluate the relevance of items → can be experts in the field, patients etc.

Item-leve CVI

- Experts are asked to rate the relevance of each item on a 4-point scale:
  - Not relevant
  - Somewhat relevant
  - Quite relevant
  - Highly relevant

Scale-level CVI

- Requires > 2 experts
- Two calculation methods

Polit DF et al. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health*. 2007;30(4):459-467

CVI example

Response scale:
☐ Not relevant (1)
☐ Somewhat relevant (2)
☐ Quite relevant (3)
☐ Highly relevant (4)

Item	Expert 1	Expert 2	Expert 3	Experts in Agreement	Item CVI
1	✓	✓	✓	3	1.00
2	✓	✓	✓	3	1.00
3	✓	✓	✓	3	1.00
4	✓	✓	✓	3	1.00
5	✓	✓	✓	3	1.00
6	✓	✓	✓	3	1.00
7	✓	✓	✓	3	1.00
8	—	✓	✓	2	.67
9	—	✓	✓	2	.67
10	✓	—	✓	2	.67
				Average I-CVI =	.90
Proportion relevant	.80	.90	1.00		

$$\text{Item-level CVI} = \frac{Number\ of\ experts\ rating\ an\ item\ 3\ or\ 4}{Number\ of\ experts}$$

$$\text{Scale-level } CVI_1 = \frac{Sum\ of\ all\ iCVI}{Total\ number\ of\ iCVI} = 0.90$$

$$\text{Scale-level } CVI_2 = \frac{Number\ items\ all\ raters\ rated\ 3\ or\ 4^{*}}{Total\ number\ of\ items} = \frac{7}{10} = 0.70$$

Commonly used criterion: 0.80

* Item 1 to 7

A small experiment

2 groups. Please follow the instructions and answer the questions according to the group you are in.

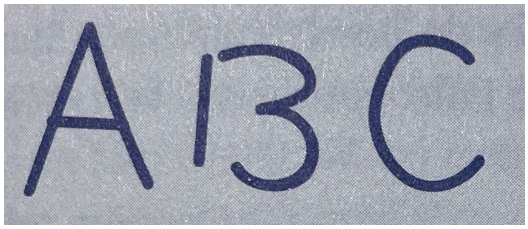
Please do **NOT** look at the other groups' questions and **DO NOT** talk to each other.

37

38

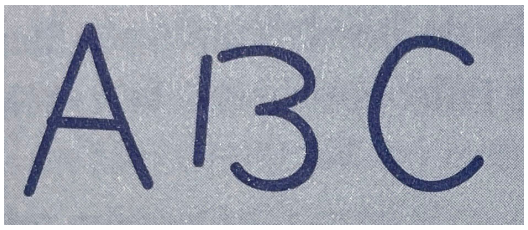
Another small experiment

Individually, write down or remember what you read? You have 3 seconds.

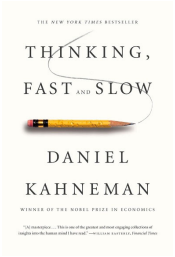


Content validity and context

Who saw what on the picture?



Content validity and context



The context of a question determines the answer

- Question order matters
- How a question is phrased matters
- Where, when and how a questionnaire is filled out matters

This is particular true when we answer questions where we are unsure about the answer

Resolved (partially) by pilot testing

*NB: Right answer is 16.344 people died from cancer in DK in 2023**

Morten Münster. Jytte fra Markeling er desværre gået for i dag

* https://www.cancer.dk/fakta-kraeft/statistik-om-kraeft/noegletal/

41

Criterion validity

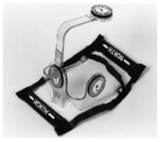
'The degree to which the scores of a measurement instrument are an adequate reflection of a gold standard'

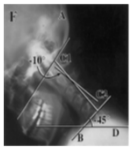
COSMIN 2011

Two types

- Concurrent validity
  - Does the instrument measure the same as a gold standard?
  - E.g. Neck function questionnaires
- Predictive validity
  - Can the instrument predict a future gold standard?
  - E.g. NDI and future neck range of motion

Gold standard: Neck range of motion





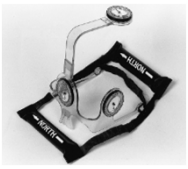
42

Criterion validity

How is this established?

- Identify suitable criterion
- Identify relevant target population
- Define *a priori* magnitude and direction of agreement
- Obtain scores for instrument and criterion
- Determine strength

↔



More on this in the advanced course…

Construct validity

Three types

Structural validity
'The degree to which the scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured'

Hypothesis testing
'The degree to which the scores of a measurement instrument are consistent with hypotheses'

Cross-cultural validity
'The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument'

COSMIN 2011

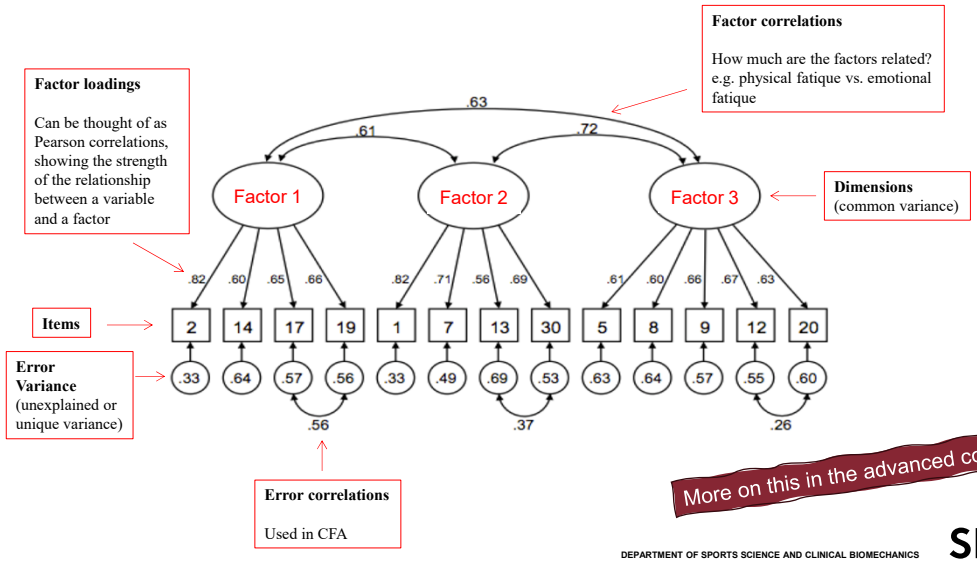
Structural validity

Factor analyses

- *Explorative FA*
  - 'Explores' the dimensionality of a questionnaire
  - Many different types
  - Used when:
    - No idea of the number of factors
    - No idea which items belong to each factor
- *Confirmatory FA*
  - 'Confirms' the dimensionality of a questionnaire
  - More rigorous statistical procedures
  - Use fit-parameters to assess whether data fit a hypothesised factor structure
  - Used when:
    - You know the number of factors (from theory or orther studies)
    - You know which items belong to each factor (from theory or other studies)

45

Structural validity



46

Hypothesis testing

"How likely is it that the questionnaire measures what it purports to measure?"

Constructing hypotheses regarding:

1. The internal structures of the instrument
2. The relationship to other instruments
3. The differences in e.g. patient groups

47

Hypothesis testing

How is this established?

- Describe the measured constructs
- Formulate expected relationships
  - Related (convergent) hypotheses
  - Non-related (divergent) hypotheses
  - Group (known group) hypotheses
- Describe the compared instruments
- Gather empirical data
- Assess results in terms of the stipulated hypotheses
- Discuss observed finding

More on this in the advanced course…

48

Cross-cultural validity

After translation and cross-cultural adaptation of an international questionnaire

- Assessing its construct validity compared to the original
- Assessment of
  - Measurement invariance (multigroup CFA)
  - Differential item functioning (logistic regression or IRT)

49

Main points – reproducibility and validity

Reproducibility
- *Internal consistency* is the interrelatedness among items measured with Cronbach's α and/or McDonald's Ω
- *Reliability* is a ratio of true variance/total variance measured with ICC and Kappa
- Both range from 0-1 → difficult to interpret

- *Measurement error* is the error on repeated measures. We use SEM, LOA and %-agreement
- Measured in same units as the scale → easy to interpret

Validity
- Content validity → item relevance, comprehensiveness and comprehensibility
- Can be measured with Content Validity Index
- Context is important
- Criterion validity uses a gold (or rather a silver) standard
- Construct validity establishes dimensionality, tests hypotheses and looks at measurement invariance
- Cross-cultural validity compared the translated version of questionnaire with the original

50

Advanced course in questionnaire technique and clinimetrics (part 2)

Dates: 9th, 16th and 23rd November 2026

Venue: AU

NB: Min. 10 participants

Course link AU: PhD Course Management (search for the course)

Aim

At the end of the course the participant will:

- Have the skills to complete the process of developing a new measurement instrument
- Have basic knowledge about item reduction and factor analysis
- Know how to perform a field test
- Be able to define, discuss and interpret the measurement properties of a) validity, b) reproducibility, c) responsiveness and d) interpretation
- Have an overview of the benefits of modern psychometric methods such as Item Response Theory (IRT) and Rasch analyses

Questions?