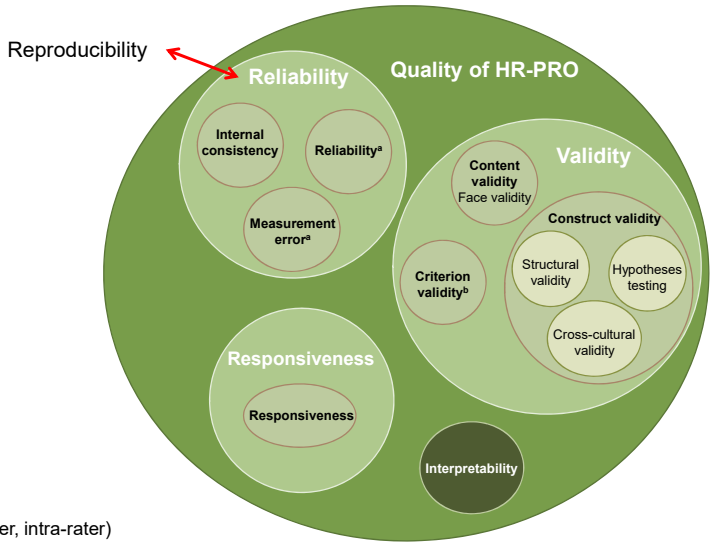
Reproducibility

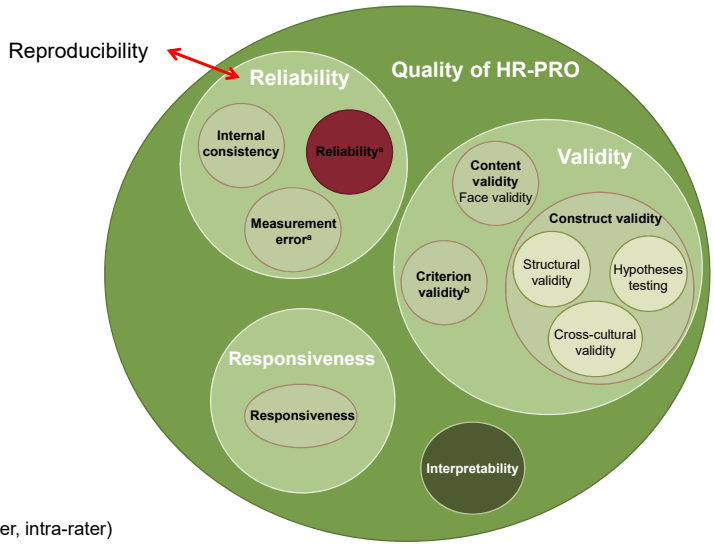
Henrik Hein Lauridsen
Associate Professor
Research unit for Clinical Biomechanics
University of Southern Denmark

The COSMIN taxonomy



[a](test–retest, inter-rater, intra-rater)
[b](concurrent validity, predictive validity)

The COSMIN taxonomy



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

[b](concurrent validity, predictive validity)

Reproducibility

Definition:

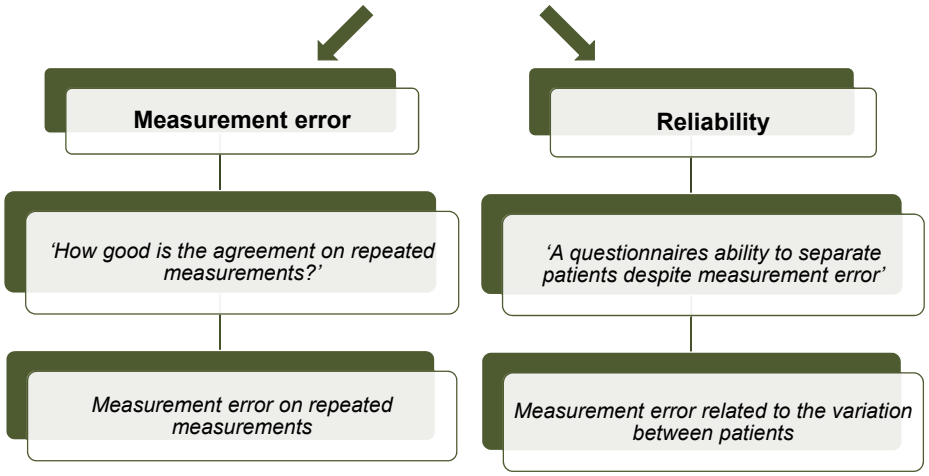
"The degree to which the measurement is free form measurement error"

"Graden af målefejl ved en given måling"

Mokkink et al. 2010

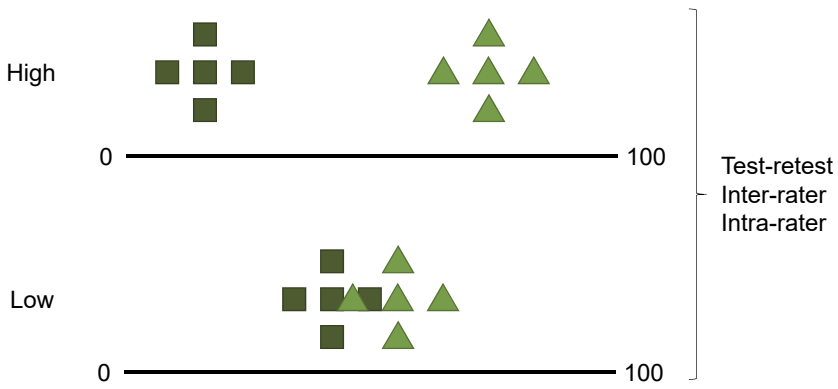
Reproducibility

'The degree to which the measurement i free from measurement error'



Reliability

5 measurements in 2 persons



Parameters

Scale	Parameters
Reliability	
Continuous	ICC
Ordinal	ICC or weighted kappa
Nominal	Unweighted kappa

- Scale from 0 -1
- Difficult to interpret

De Vet et al. "Measurement in Medicine", 2011

Classical test theory

Assumes that the true score is the sum of the observed score and random error

$$Y = \eta + \varepsilon$$

Empirical score
What we measure

=

Person characteristics
What we want to measure
True score

+

Test (and environmental) characteristics
Measurement error

If the measurement error is small we can ignore it
= *Y* is reliable

Reliability

$$Y = \eta + \varepsilon$$

Can only be disentangled when there are repeated measures
i = repeated measurement (raters, occasions etc.)

$$Y_i = \eta + \varepsilon_i$$

Assumption:
Rewritten to variance as ε_i and η are uncorrelated

$$\sigma^2(Y_i) = \sigma^2(\eta) + \sigma^2(\varepsilon_i)$$

Total variance | True variance | Error variance

Reliability

$$Y = \eta + \varepsilon$$

Can only be disentangled when there are repeated measures
i = repeated measurement (raters, occasions etc.)

$$Y_i = \eta + \varepsilon_i$$

Assumption:
Rewritten to variance as ε_i and η are uncorrelated

$$\sigma^2(Y_i) = \sigma^2(\eta) + \sigma^2(\varepsilon_i)$$

Assumption:
$\sigma^2(\varepsilon_i)$ is constant for every repetition → implies that $\sigma^2(Y_i)$ is also constant

$$\sigma^2(Y) = \sigma^2(\eta) + \sigma^2(\varepsilon)$$

$$Reliability = \frac{True\ variance}{Total\ variance}$$

Reliability

$$Y = \eta + \varepsilon$$

Can only be disentangled when there are repeated measures
i = repeated measurement (raters, occasions etc.)

$$Y_i = \eta + \varepsilon_i$$

Assumption:
Rewritten to variance as ε_i and η are uncorrelated

$$\sigma^2(Y_i) = \sigma^2(\eta) + \sigma^2(\varepsilon_i)$$

Assumption:
$\sigma^2(\varepsilon_i)$ is constant for every repetition → implies that $\sigma^2(Y_i)$ is constant

$$\sigma^2(Y) = \sigma^2(\eta) + \sigma^2(\varepsilon)$$

$\sigma^2(\eta) = \sigma_p^2;\ p = patient$

$$Reliability = \frac{True\ variance}{Total\ variance} = \frac{\sigma_p^2}{\sigma_Y^2} = \underbrace{\frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}}$$

COSMIN taxonomy

Intraclass Correlation Coefficients

ICC

- 6 different types (Shrout & Fleiss)
- 10 different types (McGraw & Wong)
- Variation ratio between patients and error (random error)
- ANOVA analyses

An example (test-retest)

10 patients measured at 2 time points

	Time points	
Patients	t_1	t_2
1	9	7
2	6	5
3	3	4
4	2	1
5	4	5
6	6	4
7	8	9
8	4	5
9	2	1
10	10	9
	5.4	5.0

Sources of variation

Interaction effect (random error) $= \sigma^2_{po,e}$

Variance between patients $= \sigma^2_p$

$\sigma^2_o =$ Variance between time points (systematic error)

	Time points	
Patients	t_1	t_2
1	9	7
2	6	5
3	3	4
4	2	1
5	4	5
6	6	4
7	8	9
8	4	5
9	2	1
10	10	9
	5.4	5.0

ICC

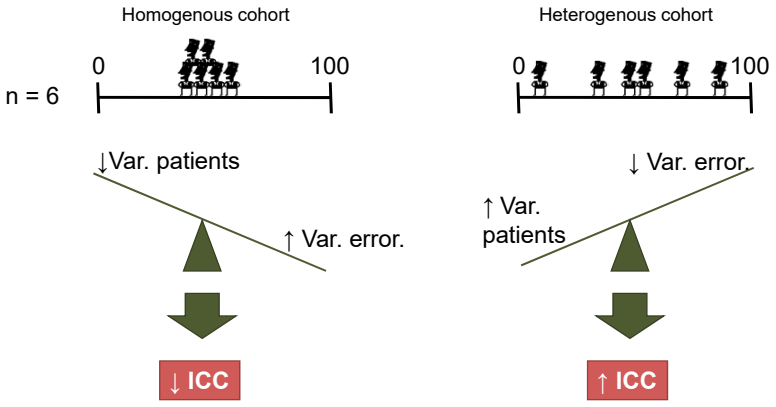
$$ICC = \frac{signal}{signal + noise} = \frac{true\ variance}{true\ variance + error\ variance}$$

$$ICC = \frac{Var_{patient}}{Var_{patient} + Var_{error}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{error}^2}$$

Shrout & Fleis, 1979; McGraw & Wong, 1996

ICC

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



Shrout & Fleis, 1979; McGraw & Wong, 1996

Types of reliability

Types	Definitions
Inter-rater reliability	Variation between 2 or more raters who measure the same group of subjects
Intra-rater reliability	Variation of data measured by 1 rater across 2 or more trials
Test-retest reliability	Variation in 2 measurements taken by an instrument on the same subject under the same conditions*

* Generally indicative of reliability in situations when raters are not involved or rater effect is neglectable, such as self-report survey instruments

Choice of ICC formula

Depends on:

I. Model
- One-way or two-way ANOVA
- Inclusion of 'Random' or 'fixed' factors

II. Definition of relationship which is important
- Ranking ('consistency') or absolute agreement ('agreement')

III. Type
- Single rating or means of multiple ratings

Model

One-way ANOVA
- When only interested in distinguishing between subjects and not in other factor(s).

Two way ANOVA
- When also interested in other factors, e.g. subjects AND time-points (2 factors)

Example: one- or two-way

One-way ANOVA
- Variation in birth weight of twins

Two-way ANOVA
- Is birth weight of *first*-born baby higher than *second*-born in twins

Model

Random factors

- One wants to generalise to all other representors of that factor (observers or moments)

Fixed factors

- One is only interested in the representors of the factor under study (observers or moments)

Example: random or fixed factors

Random factors

- How reliable are the judgements of X-rays by an ‘average’ radiologist?

Fixed factors

- How reliable are the judgements of the two radiologists who rate the X-rays in this study?

Definition of relationship which is important

Consistency
- Only random error is included – ranking is important

Agreement
- Systematic errors are also included – absolute agreement is important

Type

Single rating
- Comparing 1 rating from each of e.g. 4 raters

Multiple ratings
- Comparing the mean rating of e.g. 4 raters

If the mean value is taken of multiple ratings, the error variance becomes smaller as random errors diminish by averaging

Data on individual ratings more common

ICC forms

McGraw and Wong (1996)				Shrout and Fleiss (1979)
Model		***Relationship***	***Type***	
One-way	Random effects	Agreement	Single rater/measurement	ICC (1,1)

Model

Type

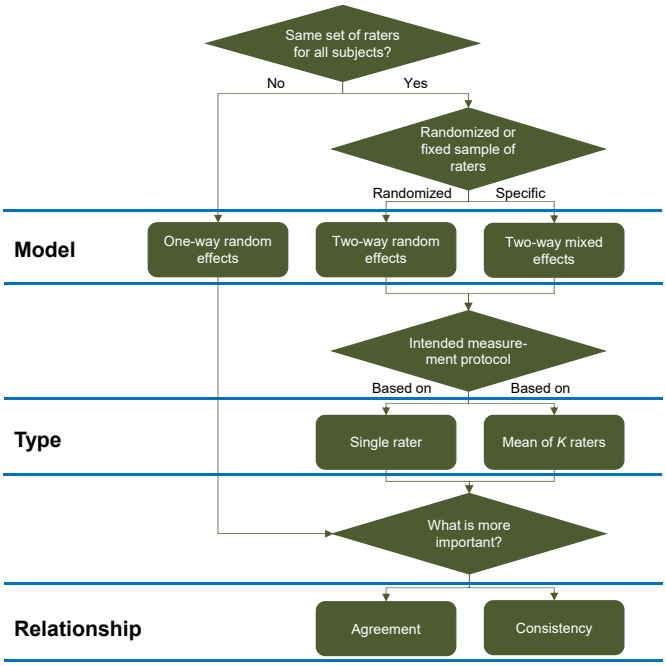
ICC forms

McGraw and Wong (1996)				Shrout and Fleiss (1979)
Model		Relationship	Type	
One-way	Random effects	Agreement	Single rating/measurement	ICC (1,1)
One-way	Random effects	Agreement	Multiple ratings/measurements	ICC (1,K)
Two-way	Random effects	Agreement	Single rating/measurement	ICC (2,1)
Two-way	Random effects	Agreement	Multiple ratings/measurements	ICC (2,K)
Two-way	Random effects	Consistency	Single rating/measurement	-
Two-way	Random effects	Consistency	Multiple ratings/measurements	-
Two-way	Mixed effects	Agreement	Single rating/measurement	-
Two-way	Mixed effects	Agreement	Multiple ratings/measurements	-
Two-way	Mixed effects	Consistency	Single rating/measurement	ICC (3,1)
Two-way	Mixed effects	Consistency	Multiple ratings/measurements	ICC (3,K)

Choice depends on study design: inter-rater, intra-rater or test-retest

Koo et al., 2016

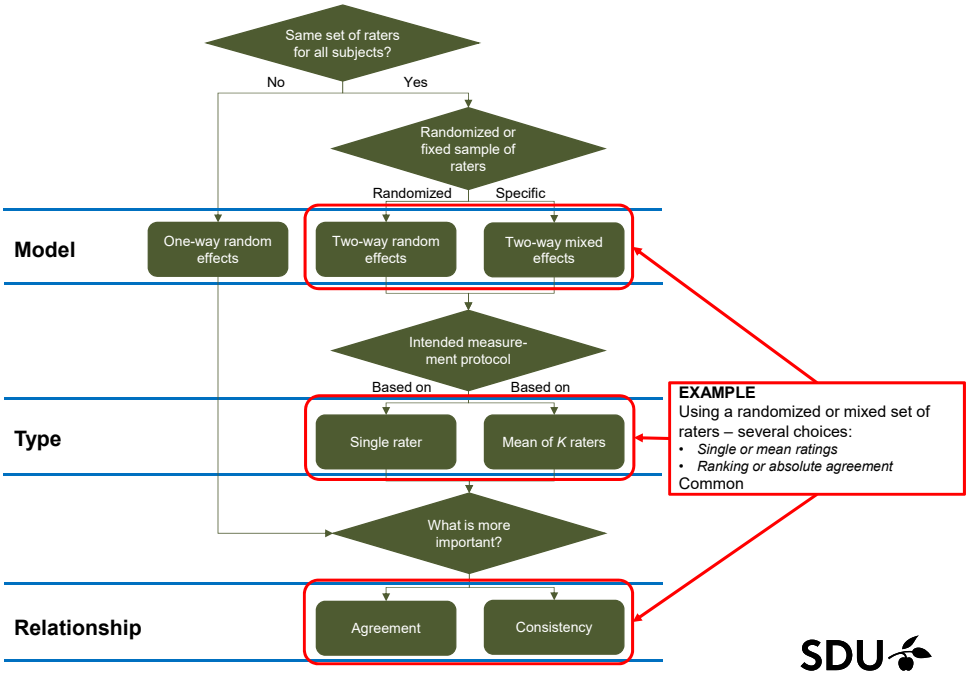
Flowchart



SDU

Flowchart

Inter-rater reliability



SDU

ICC (2,1)$_{consistency}$

$$ICC(2,1)_{consistency} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{po,e}^2}$$

Random error (residual)

- Two-way random-effects ANOVA
- Not including systematic error
- Used in analyses where the systematic error is known to be small

Shrout & Fleis 1979; McGraw & Wong 1996, Weir 2005

ICC (2,1)$_{agreement}$

$$ICC(2,1)_{agreement} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2 + \sigma_{po,e}^2}$$

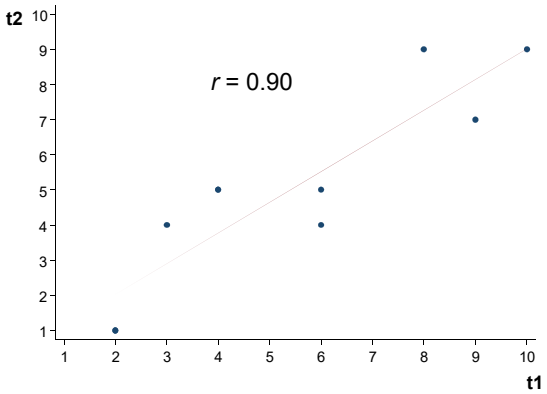
Systematic error between t1 and t2

Random error (residual)

Two-way random-effects ANOVA
Includes systematic error
Often used in test-retest designs

Shrout & Fleis 1979; McGraw & Wong 1996, Weir 2005

What about Pearson's *r*?

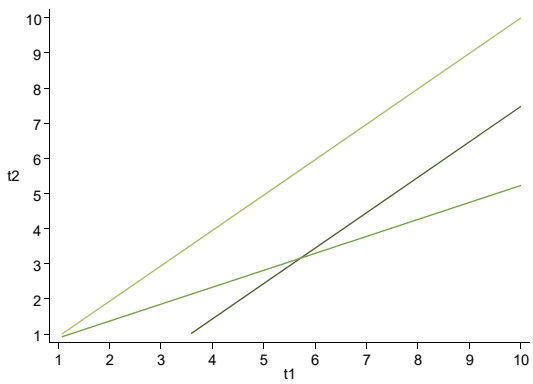


Patienter	Measurements t_1	t_2
1	9	7
2	6	5
3	3	4
4	2	1
5	4	5
6	6	4
7	8	9
8	4	5
9	2	1
10	10	9
	5.4	5.0

Does not consider:

- Systematic error
- Variance differences between t1 and t2

Pearson vs. ICC



$ICC_{a,c} = 1, r = 1$

$ICC_a \neq 1, ICC_c = 1, r = 1$

$ICC_{a,c} \neq 1, r = 1$

Kappa

"A measure of reliability between two measurements which is adjusted for chance agreement."

Nominal scale: kappa
Ordinal scale: weighted kappa

Nominal scale - example

		Radiologist B		
		+	-	Total
Radiologist A	+	50	15	*65*
	-	15	20	*35*
	Total	*65*	*35*	*100*

Observed agreement (p_o): (50 + 20)/100 = 0.70

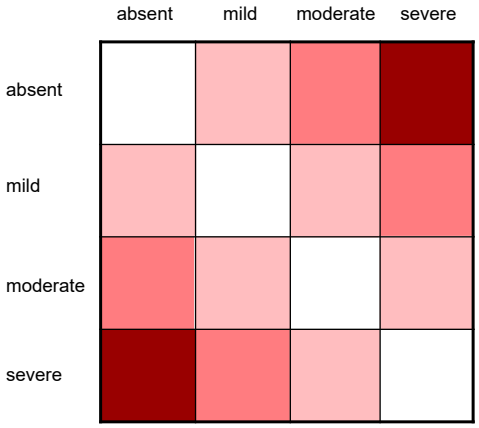
Expected agreement (p_e) = 0.54

$\kappa = (p_o - p_e) / (1 - p_e)$

κ = (0.70-0.54) / (1-0.54) = 0.35

Ordinal scale - weighted kappa

E.g. Classification of malignant melanoma



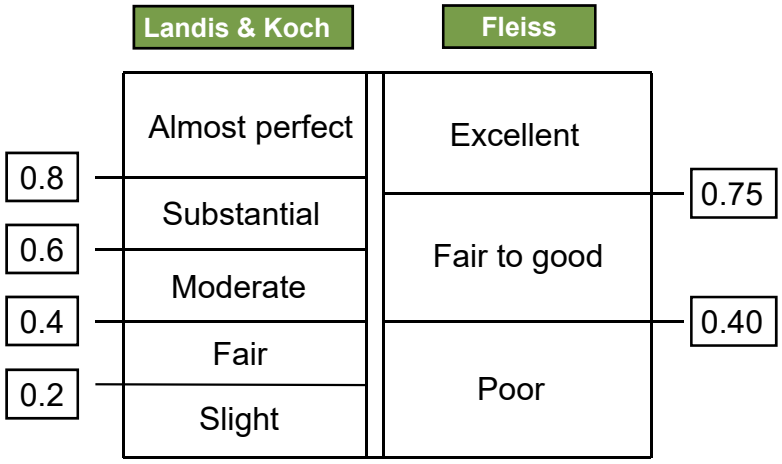
Some ‘errors’ are weighted more than others

Interpretation

Kappa depends on:

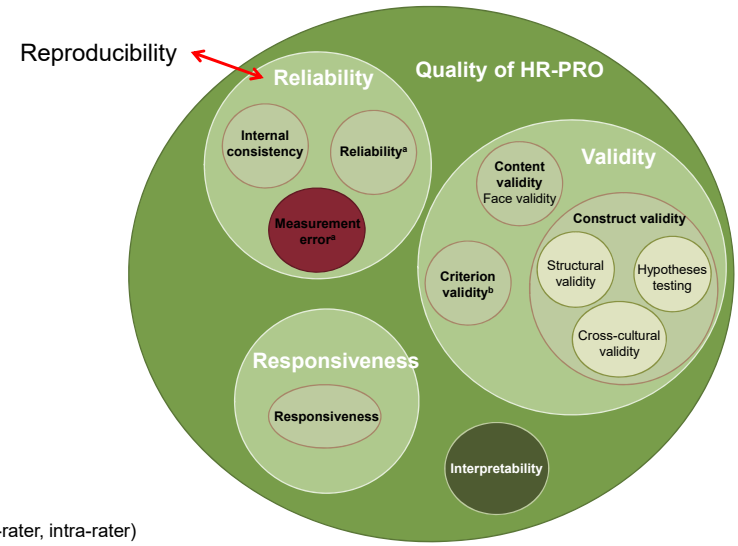
1. The number of response options
   - More response options → lower kappa
2. The score prevalence
   - Skewed distributions → lower kappa
3. Presence of systematic differences
   - Slightly higher kappa

Interpretation



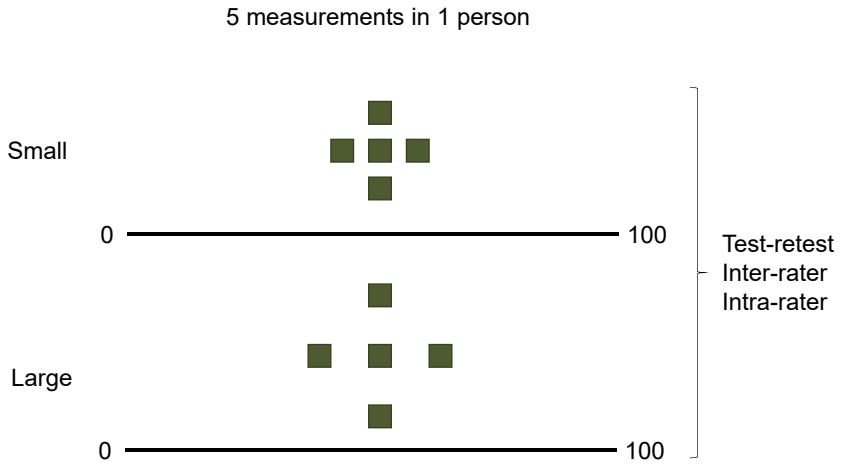
Landis and Koch, the measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174
Fleiss JL. Statistical methods for rates and proportions, 2nd ed. New York, John Wiley & Sons, 1981.

The COSMIN taxonomy



[a](test–retest, inter-rater, intra-rater)
[b](concurrent validity, predictive validity)

Measurement error



Measurement error

Scale	Parameters
Continuous	Standard error of the measurement (SEM) Limits of agreement (LOA)
Ordinal	% agreement
Nominal	% agreement

- Same units as the measurement instrument
- Easy to interpret
- A characteristic of the questionnaire

De Vet et al. "Measurement in Medicine", 2011

Standard error of the measurement (SEM)

Definition

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

Purpose

- Quantify the extend to which a 'test' provides accurate scores
  - ↓ SEM → ↑ score accuracy and vice versa

Use

- SEM used to calculate confidence intervals around obtained scores
- E.g.
  - 68% CI = Score ± (1*SEM)
  - 95% CI = Score ± (1.96*SEM)
  - 99% CI = Score ± (2.58*SEM)

Example

- Depression score = 50 (score range [0-100])
- How confident are we that the person's true depression score is 50?

De Vet et al. 2006

SEM

$$SEM = SD\left(\sqrt{1 - reliability}\right)$$

Example

- Depression score with SD = 15 (cross-sectional design)
- What is SEM of a single depression score = 50 under three different conditions:

- Reliability = 0.70: $SEM = 15\left(\sqrt{1 - 0.7}\right) = 8.22$ 95% CI = 50 ± (1.96 x 8.22) = [66.11; 33.89]
- Reliability = 0.80: $SEM = 15\left(\sqrt{1 - 0.8}\right) = 6.71$ 95% CI = 50 ± (1.96 x 6.71) = [63.15; 36.85]
- Reliability = 0.90: $SEM = 15\left(\sqrt{1 - 0.9}\right) = 4.74$ 95% CI = 50 ± (1.96 x 4.74) = [59.29; 40.71]

This is the reason single scores have to be more reliable → usually > 0.9
In group scores, the measurement error ↓ by a factor √n → usually > 0.7

de Vet et al. Measurement in Medicine, p 142.

SEM

Longitudinal design

- SEM used to calculate the measurement error of an instrument over time
- Established in a test-retest design on stable patients
- Can be extracted from the ICC when all variance components are available

$SEM_{consistency}$

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

$$Standard\ Error\ of\ Measurement\ (SEM_{consistency}) = \sqrt{\sigma^2_{error}}$$

$$SEM_{consistency} = SD \times \sqrt{(1 - ICC_{consistency})} \qquad (\mathrm{SD} = SD_{pooled} = (SD_{time1} + SD_{time2})/2)$$

$$SEM_{consistency} = SD_{change}/\sqrt{2}$$

$SEM_{agreement}$

$$Reliability = ICC_{agreement} = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma^2_{measurements} + \sigma^2_{error}}$$

$$Standard\ Error\ of\ Measurement\ (SEM_{agreement}) = \sqrt{\sigma^2_{measurements} + \sigma^2_{error}}$$

$$SEM_{agreement} = SD \times \sqrt{(1 - ICC_{agreement})} \qquad (\mathrm{SD} = SD_{pooled} = (SD_{time1} + SD_{time2})/2)$$

Bland & Altman LOA

Graphical presentation of mean and difference between two scores

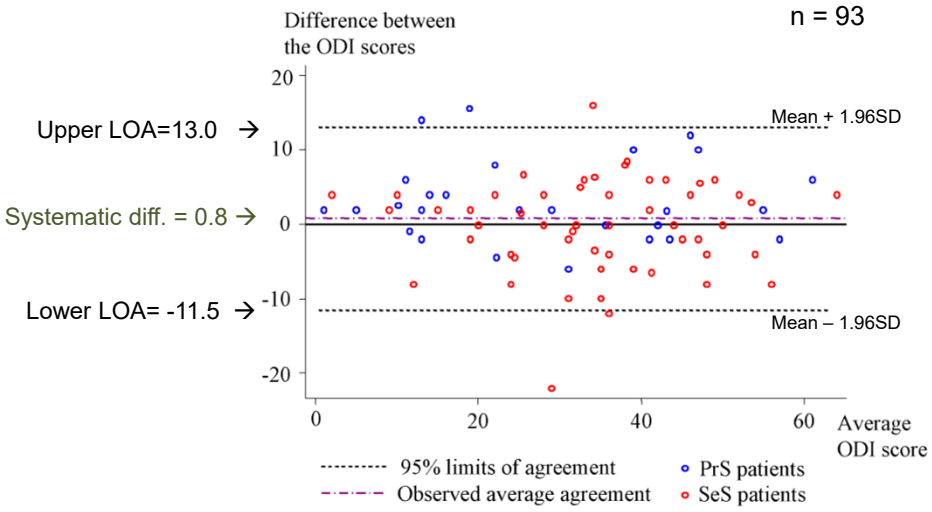
- Limits of agreement (LOA) = measurement error

$$LOA = mean_{change} \pm 1.96 \cdot SD_{change}$$

Systematic error — Random error

95% of the differences lies between the LOA

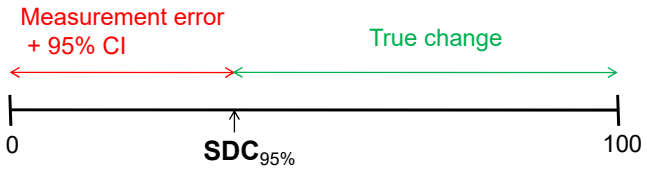
Bland & Altman plot (agreement)



Smallest detectable change (SDC)

Is the smallest change in score that you **CAN** detect with the instrument, above measurement error (in individual patients)

$$SDC_{agreement} = 1.96 \cdot SEM_{agreement} \times \sqrt{2}$$



NB: Also called Minimal Detectable Change, Real change, Smallest Real Change, Significant change

SEM, LOA and SDC

SEM

$$SEM_{consistency} = SD_{change}/\sqrt{2}$$

Limits of agreement (LOA)

$$LOA = mean_{difference} \pm 1.96 \times SD_{change}$$

$$LOA = mean_{difference} \pm 1.96 \times SEM_{consistency} \times \sqrt{2}$$

Smallest detectable change ($SDC_{95\%}$)

$$SDC_{consistency} = 1.96 \times SD_{change}$$

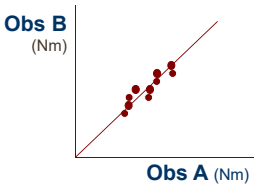
$$SDC_{consistency} = 1.96 \times SEM_{consistency} \times \sqrt{2}$$

$$SDC_{consistency} = 1.96 \times SD \times \sqrt{(1 - ICC_{consistency})} \times \sqrt{2} = 1.96 \times SD \sqrt{2 \times (1 - ICC_{consistency})}$$

$$SDC_{agreement} = 1.96 \times SEM_{agreement} \times \sqrt{2}$$

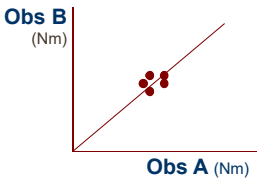
$$SDC_{agreement} = 1.96 \times SD \sqrt{(1 - ICC_{agreement})} \times \sqrt{2} = 1.96 \times SD \sqrt{2 \times (1 - ICC_{agreement})}$$

Example: What is the reliability (Rel) and measurement error (SEM)?



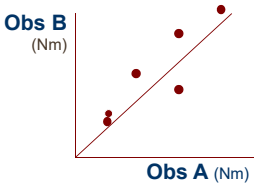
Rel = 0.986

SEM = 1.48



Rel = 0.5

SEM = 1.48



Rel = 0.993

SEM = 4.39

Stratford et al, 1989

Summary

Reproducibility

- Reliability
  - Measurement error related to the variation between patients
  - Discriminative instruments
  - Parameters: ICC, Kappa
- Measurement error
  - Measurement error on repeated measurements
  - Evaluative instruments
  - Parameters: SEM, LOA (SDC)

SDU