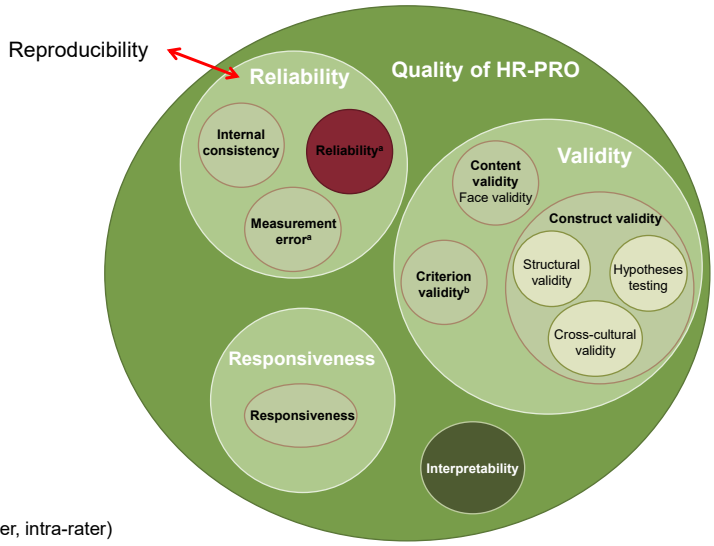
Reproducibility

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



1

The COSMIN taxonomy



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

2

Reproducibility

Definition:

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

"Graden af målefejl ved en given måling"

N.B. A brief recap…

Mokkink et al. 2010

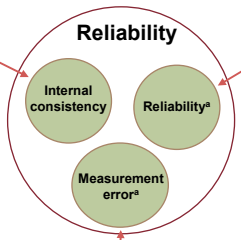
3

COSMIN 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

4

Examples

Measurement error (= absolute agreement)

- How much do repeated measurements differ when no real change has occurred?
- Expressed in original units (e.g. kilo)

Example

- Two raters weigh the same person, with no real change in weight. How much do their readings differ?

Reliability (= relative reliability)

- How well can we distinguish between people despite measurement error?
- ↓ measurement error compared to sample variation → ↑ reliability
- Reported as a unitless coefficient from 0 – 1

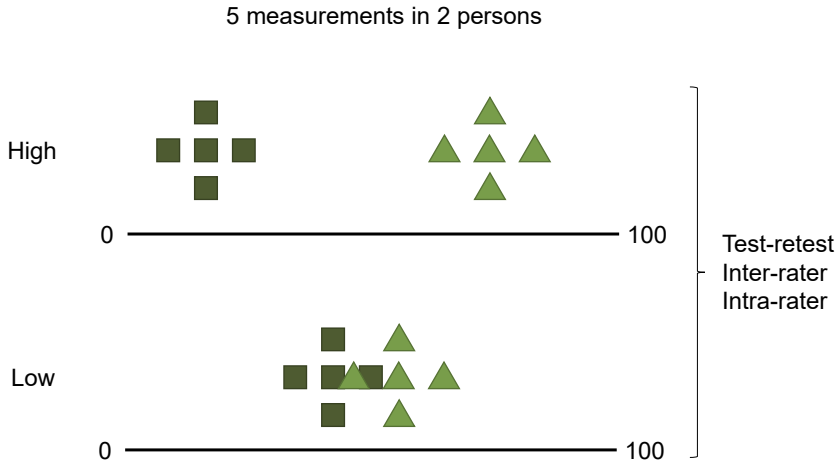
Example

- Assume the same measurement error (0.5 kg) in both groups
- Adult body weight variation: 50 – 120 kg (↑ *hetero*geneity) → ↑ reliability
- Baby body weight variation: 1.5 – 5.0 kg (↑ *homo*geneity) → ↓ reliability

de Vet HC, Mokkink LB, Terwee CB, *et al.* Clinicians are right not to like Cohen's kappa. *BMJ* 2013;346:f2125.

5

Reliability



6

Parameters

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

- Scale from 0 -1
- Difficult to interpret

* ICC(2.1)
** ICC(2.1) if the ordinal score is treated as approximately continuous
De Vet et al. "Measurement in Medicine", 2011

7

Classical test theory

Assumes that the measured score is the sum of the true 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

8

Reliability

See Measurement in Medicine

$$Y = \eta + \varepsilon$$

η and ε can only be disentangled when there are repeated measures
i = repeated measurements (raters, occasions etc.)
True score η is assumed stable across repetitions

$$Y_i = \eta + \varepsilon_i$$

Variance decomposition:
Rewritten to variance as η and ε_i are uncorrelated

$$variance(Y_i) = variance(\eta) + variance(\varepsilon_i)$$

Renaming using statistical symbols for variance

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

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

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

Apply COSMIN definition of reliability:
The proportion of the total variance, which is due to the 'true' difference between the (e.g.) patients

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

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

Intraclass Correlation Coefficients

ICC

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

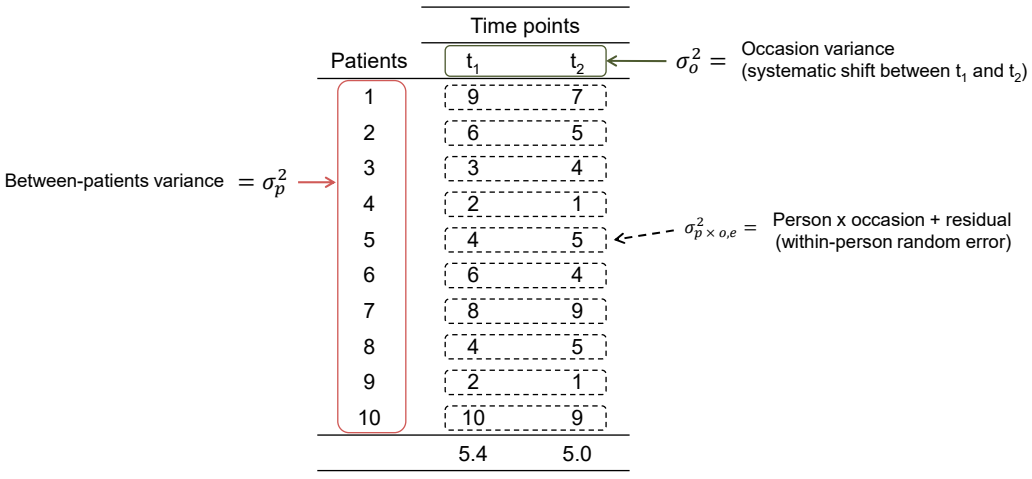
An example (test-retest)

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

11

Sources of variation



12

ICC

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

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

$$Consistency = \sigma_{error}^2 = \sigma_{po,e}^2$$

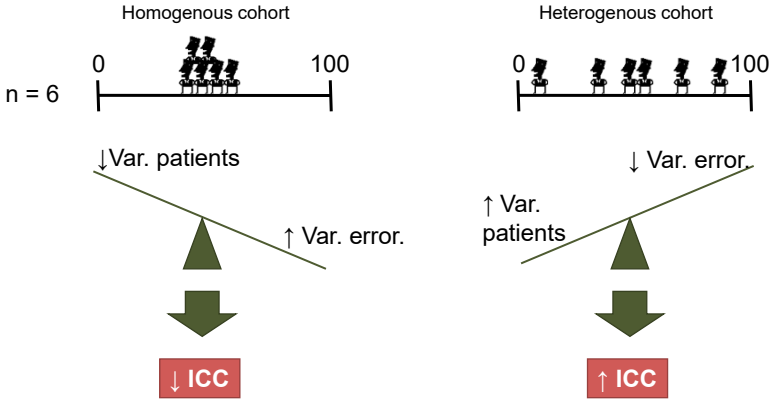
$$Agreement = \sigma_{error}^2 = \sigma_o^2 + \sigma_{po,e}^2$$

Shrout & Fleis, 1979; McGraw & Wong, 1996

13

ICC

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



Shrout & Fleis, 1979; McGraw & Wong, 1996

14

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. Effects model (ANOVA)
- One-way random-effects model
- Two-way random-effects model
- Two-way mixed-effects model

II. Unit of measurement
- Single measurement
- Average of k measurements

III. Agreement type
- Consistency
- Agreement

Effects model: one-way versus two-way

One-way effects model
- Variation in birth weight of twins
- Variation of scores between patients



Two-way effects model
- Is birth weight of *first*-born baby higher than *second*-born in twins
- Is the score of patients different from *time 1 to time 2*

Effects model: random versus mixed

Random models
- How reliable are the measurements of twin birth weight of *randomly chosen* clinicians? (inter-rater)
- How reliable would patient scores be if measured again at a *typical* future occasion? (test-retest)

Mixed models
- How reliable are the measurements of twin birth weight by *the two clinicians* in this study? (inter-rater)
- How reliable are the scores of patients at these *two exact time points* in this study? (test-retest)

(i.e., babies/patients = random factor; rater/occasion = fixed factor)

Unit of measurement

Single measurement

- Used when decisions are based on one observation*

Average of *k* measurement

- Used when practice is to average k replicates (e.g., 3 trials per session)

Averaging k independent replicates reduces error variance to $\frac{\sigma_E^2}{k}$

Most studies report single-measure reliability - state clearly if you report average-measurements

* Test-retest, inter- or intra-rater reliability

Agreement type

Consistency

- Ignores systematic shift - ranking is important
- Use when you care that subjects keep the *same order*, and *systematic shift is negligible*

Absolute agreement

- Systematic shift are also included – absolute agreement is important

ICC forms - example

Model		Relationship	Type	McGraw & Wong notation (1996)	Shrout & Fleiss notation (1979)
Two-way	Random effects	Agreement	Single rater/measurement	ICC(A,1)	ICC (1,1)

Is birth weight of first-born baby higher than second-born in twins

Reliability of birth weight in twins when using randomly chosen clinicians

Including systematic shift

Single rating of birth weight in twins

Relationship

Type

Model

Type

ICC forms

= same ICC point-estimate formula; model interpretation may differ

Model		Relationship	Type	McGraw & Wong notation (1996)	Shrout & Fleiss notation (1979)
One-way	Random effects	Agreement	Single rating/measurement	ICC(1)	ICC (1,1)
One-way	Random effects	Agreement	Average of k ratings/measurements	ICC(k)	ICC (1,k)
Two-way	Random effects	Agreement	Single rating/measurement	ICC(A,1)	ICC (2,1)
Two-way	Random effects	Agreement	Average of k ratings/measurements	ICC(A,k)	ICC (2,k)
Two-way	Random effects	Consistency	Single rating/measurement	ICC(C,1)	-
Two-way	Random effects	Consistency	Average of k ratings/measurements	ICC(C,K)	-
Two-way	Mixed effects	Agreement	Single rating/measurement	ICC(A,1)	-
Two-way	Mixed effects*	Agreement	Average of k ratings/measurements	ICC(A,k)	-
Two-way	Mixed effects	Consistency	Single rating/measurement	ICC(C,1)	ICC (3,1)
Two-way	Mixed effects*	Consistency	Average of k ratings/measurements	ICC(C,k)*	ICC (3,k)*

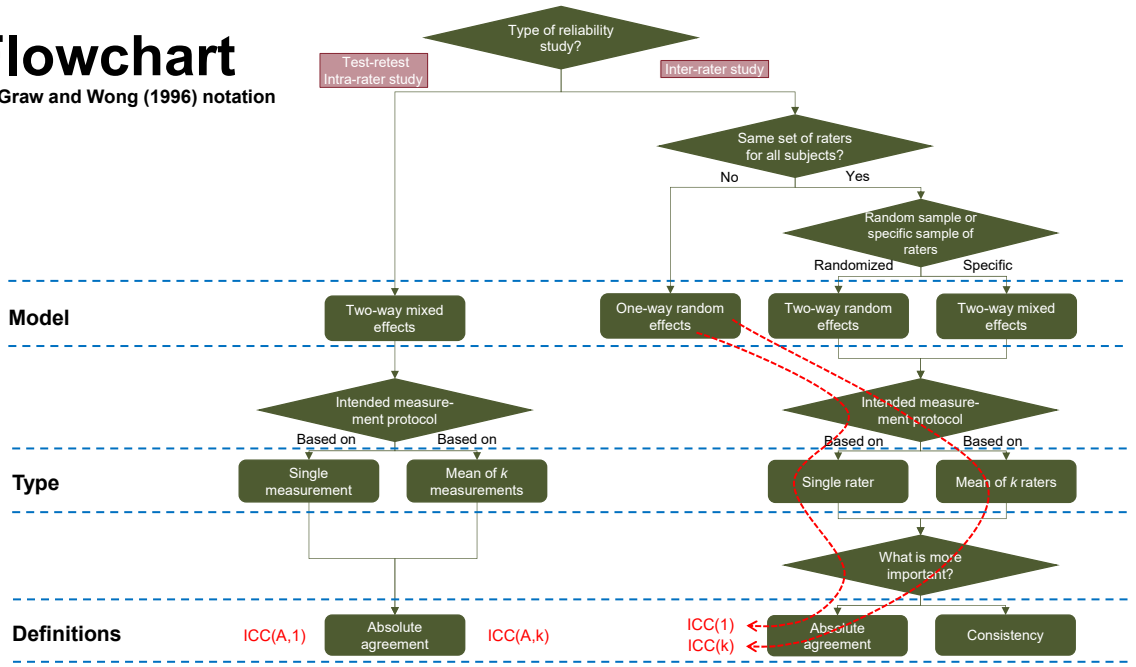
*According to McGraw and Wong (1996), the average-measure ICC for two-way mixed-effects models is estimable under Case 3A (without interaction), but not under Case 3 (with interaction). Under Case 3A, ICC(C,k) and ICC(3,k) have the same calculation formula.

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

Qin et al., 2019; Koo et al., 2016

Flowchart

McGraw and Wong (1996) notation

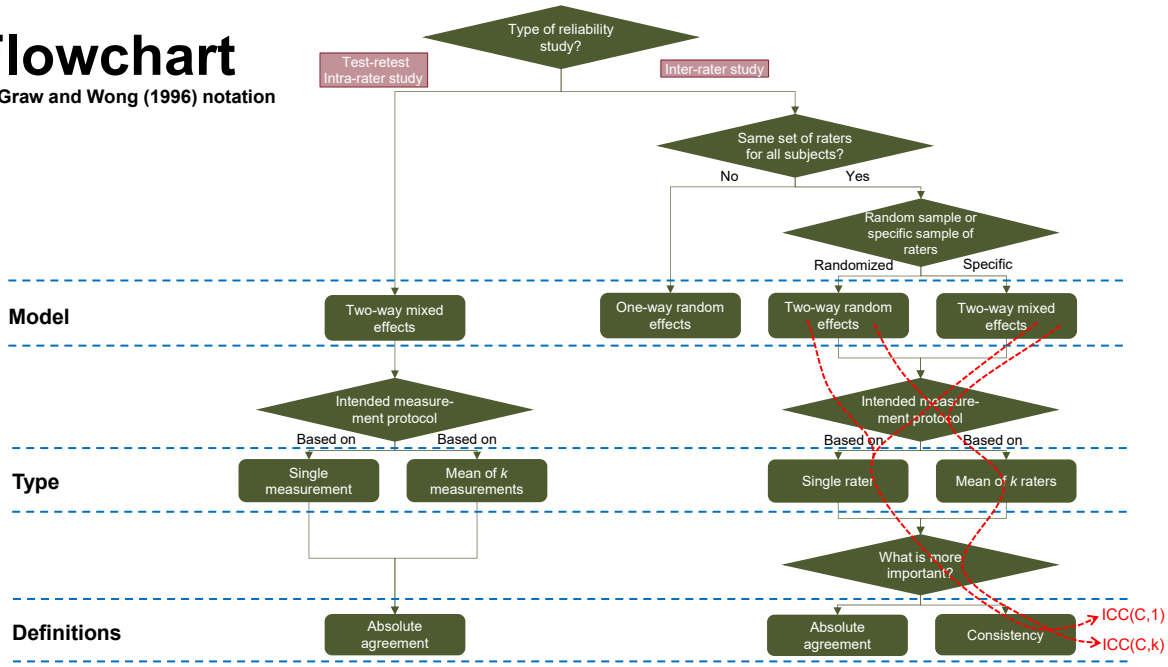


Qin et al., 2019; Koo et al., 2016

23

Flowchart

McGraw and Wong (1996) notation

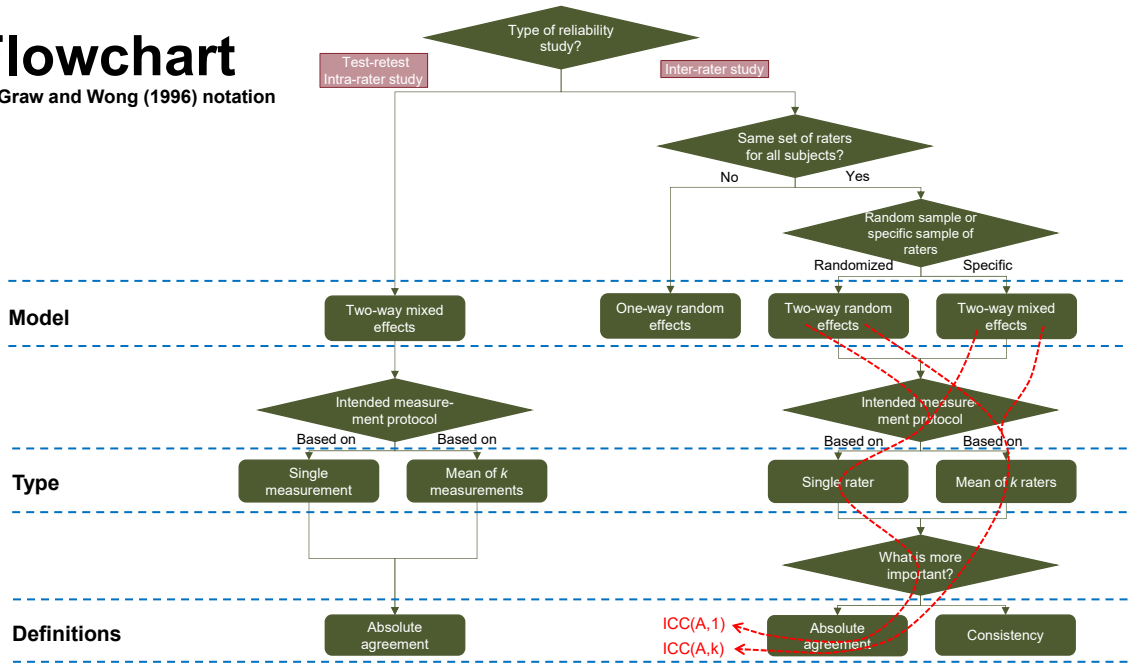


Qin et al., 2019; Koo et al., 2016

24

Flowchart

McGraw and Wong (1996) notation



Qin et al., 2019; Koo et al., 2016

25

ICC (C,1): Consistency

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

Random error (residual)

- Two-way mixed-effects ANOVA (fixed observers)
- Excludes systematic differences between observers (or measurement occasions)

McGraw & Wong 1996

26

ICC (A,1): Absolute agreement

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

Systematic error between measurement occasions

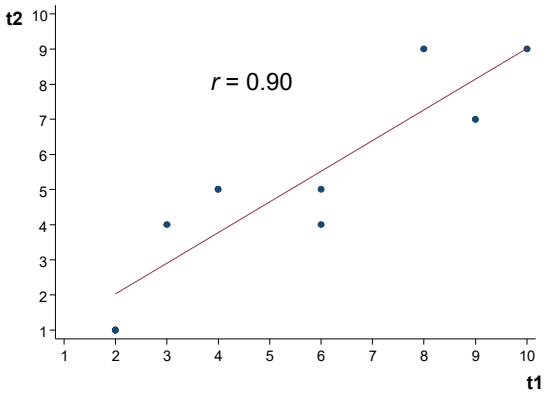
Random error (residual)

- Two-way mixed effects ANOVA
- Includes systematic differences between observers (or measurement occasions)

McGraw & Wong 1996

27

What about Pearson's *r*?



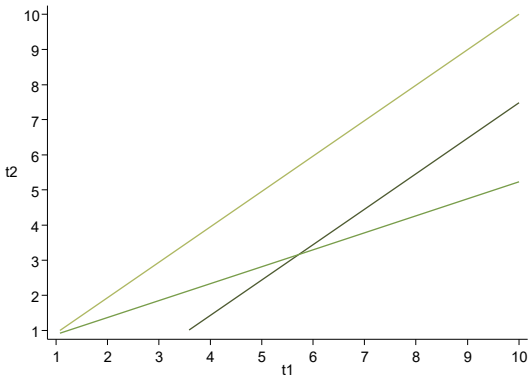
Patienter	Measurements t_1	t_2
1	9	7
2	6	5
3	3	4
4	2	1
5	4	5
6	6	4
7	8	9
8	4	5
9	2	1
10	10	9
	5.4	5.0

Does not consider:

- Systematic error
- Variance differences between t1 & t2

28

Pearson's *r* vs. ICC



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

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

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

$ICC_{consistency}$ ignores additive bias but not multiplicative bias

29

Parameters

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

- Scale from 0 -1
- Difficult to interpret

* ICC(2.1)
** ICC(2.1) if the ordinal score is treated as approximately continuous
De Vet et al. "Measurement in Medicine", 2011

30

Kappa

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

Nominal scale: kappa
Ordinal scale: weighted kappa

Is a measure of *relative* reliability

Can answer questions like

- Can a new classification system better distinguish between two categories compared to an old system?
- Can we identify patients with abnormalities in a large population sample?

Nominal scale - example

Cancer test A	Cancer test B: +	Cancer test B: -	Total
+	50	25	*65*
-	5	20	*35*
Total	*65*	*35*	*100*

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

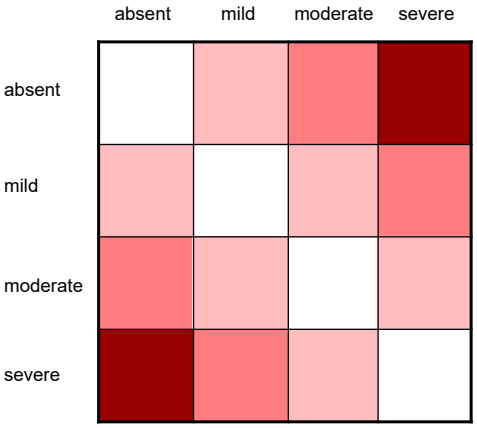
Expected agreement (p_e) = 0.53

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

$$\kappa = (0.70 - 0.53) / (1 - 0.53) = 0.37$$

Ordinal scale - weighted kappa

E.g. Classification of malignant melanoma



Some 'errors' are weighted more than others

33

Interpretation

Skewed distributions → lower kappa

Kappa depends on:

1. The number of response options
   - More response options → lower kappa

2. The score prevalence
   - Kappa is influenced by the prevalence of the attribute
   - Exists when the proportion of agreements on the positive classification (cell a) differs from that of the negative classification (cells d)
   - Measured by 'The prevalence index' (range [0-1])
   - Skewed distributions → lower kappa

		Cancer test B		
		+	-	Total
Cancer test A	+	[a] 50	[b] 25	*65*
	-	[c] 5	[d] 20	*35*
	Total	*65*	*35*	*100*

$$Prevalence\ index = \frac{|a-d|}{n} = \frac{|50-20|}{100} = 0.3$$

		Cancer test B		
		+	-	Total
Cancer test A	+	[a] 38	[b] 15	*53*
	-	[c] 15	[d] 32	*47*
	Total	*53*	*47*	*100*

$$Prevalence\ index = \frac{|a-d|}{n} = \frac{|38-32|}{100} = 0.06$$

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy* 2005;85(3):257–68.

34

Interpretation

		Cancer test B		
		+	-	Total
Cancer test A	+	[a] 50	[b] 25	*65*
	-	[c] 5	[d] 20	*35*
	Total	*65*	*35*	*100*

$$Bias\ index = \frac{|b-c|}{n} = \frac{|25-5|}{100} = 0.2$$

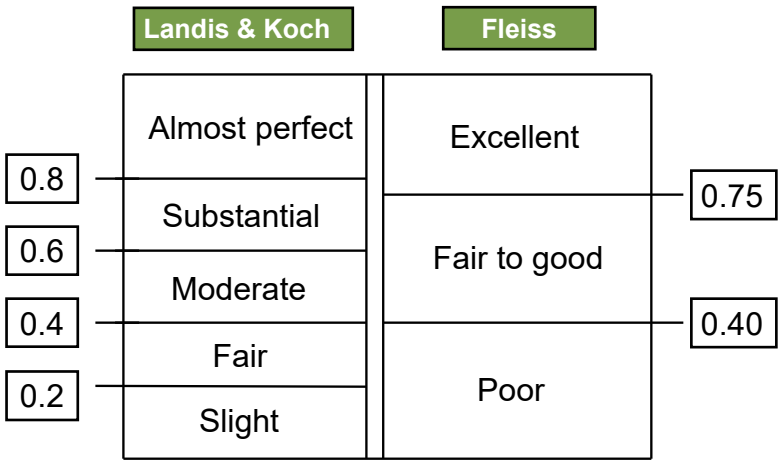
The effect of the bias index on kappa is complex.
A higher bias index can increase kappa when observed agreement and prevalence index are held constant.

Kappa depends on:

3. The rater disagreement on the proportion of positive (or negative) cases
   - Measured by the 'Bias index'
   - Reflects the extent to which disagreements are asymmetric between the discordant cells b and c
   - Bias index = 0 when both raters classify the same proportion as positive

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy* 2005;85(3):257–68.

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 Kappa paradox

The finding of low kappa values despite high percentage of agreement

Causes

- Skewed prevalence
- (Bias)

Solutions

- PABAK – prevalence-adjusted bias-adjusted Kappa
- Gwet's AC1

37

PABAK: prevalence-adjusted bias-adjusted kappa

Stata: kappaect → use Brennan & Prediger; R: use package irrCAC

PABAK

- PABAK removes the effects of prevalence and bias on κ
- Equivalent to calculating κ after setting expected chance agreement to 0.50
- For binary data: $PABAK = 2 \cdot P_o - 1$

Observed agreement

Cancer test A		Cancer test B: +	Cancer test B: -	Total
	+	[a] 50	[b] 25	*65*
	-	[c] 5	[d] 20	*35*
	Total	*65*	*35*	*100*

$$Prevalence\ index = \frac{|50 - 20|}{100} = 0.3$$

$$\kappa = 0.34$$

Cancer test A		Cancer test B: +	Cancer test B: -	Total
	+	[a] 35	[b] 15	*50*
	-	[c] 15	[d] 35	*50*
	Total	*50*	*50*	*100*

$$P_0 = (a + d)/N_o = \frac{35 + 35}{100} = 0.70$$

$$PABAK = 2 \cdot P_o - 1 = (2 \cdot 0.70) - 1 = 0.40$$

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy. 2005 Mar;85(3):257–68.

38

Gwet's AC 1 – an alternative to PABAK

- PABAK fixes expected chance agreement at **0.50** for binary ratings
- AC1 instead estimates chance agreement from the observed category frequencies
- Unlike Cohen's kappa, the AC1 chance estimate does not become large simply because one category dominates
- AC1 is therefore less affected by the prevalence problem while still reflecting the observed rating distribution

In simple terms:

- PABAK imposes a 50/50 category distribution
- AC1 uses the observed data, but applies a different, more stable estimate of chance agreement

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy. 2005 Mar;85(3):257–68.

39

Gwet's AC1

AC1 estimates less agreement as being due to chance than Cohen's kappa
Found to be 'paradox-resistant' in case of unstable Kappas

		Cancer test B +	Cancer test B -	Total
Cancer test A	+	[a] 50	[b] 25	*65*
	-	[c] 5	[d] 20	*35*
	Total	*65*	*35*	*100*

$$Prevalence\ index = \frac{|50 - 20|}{100} = 0.3$$

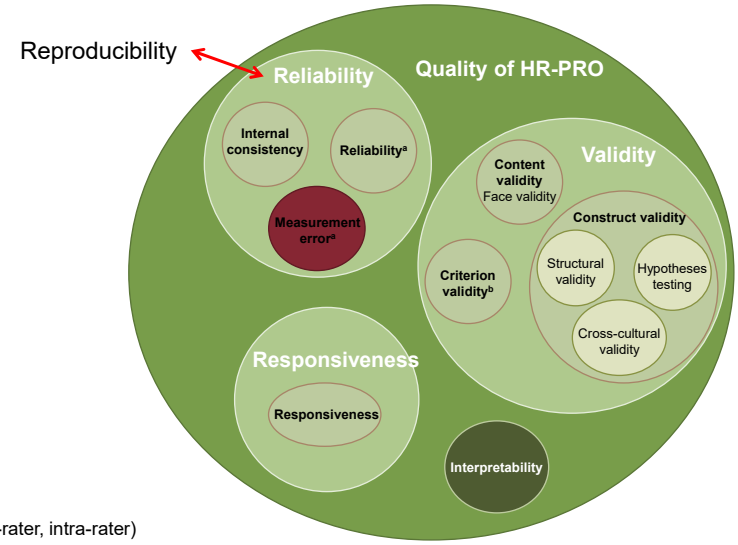
$$Bias\ index = \frac{|25 - 5|}{100} = 0.2$$

Statistic	Value
Observed agreement	0.70
Cohen's κ	0.37
PABAK	0.40
AC1	0.45

Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013 Apr 29;13:61.

40

The COSMIN taxonomy

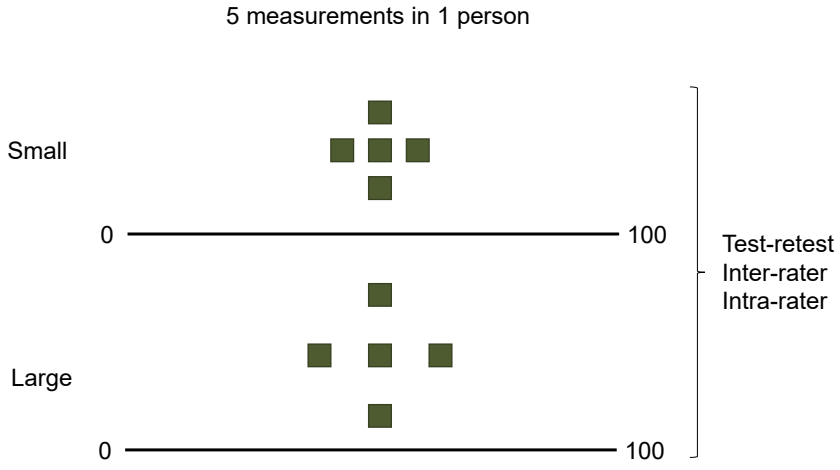


[a](test–retest, inter-rater, intra-rater)

[b](concurrent validity, predictive validity)

41

Measurement error



42

Measurement error

Scale	Parameters
Continuous	Standard Error of the Measurement (SEM) Bland & Altman - Limits of Agreement (LOA)
Ordinal	Proportion of specific agreement
Nominal	Proportion of specific agreement

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

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

Proportion of specific agreement

Expresses agreement separately for positive and negative ratings, providing *a clearer and more clinically meaningful* assessment than general agreement

First reported by Dice in 1945 but NOT found broad application

A rheumatology example

- Question: Will my diagnosis of progression or not agree with a colleague's diagnosis?

		Rheumatologist A		
		Progression	No progression	Total
Rheumatologist B	Progression	[a] 33	[b] 4	37
	No progression	[c] 4	[d] 5	9
	Total	37	9	46 (N)

de Vet HC, Mokkink LB, Terwee CB, *et al*. Clinicians are right not to like Cohen's kappa. *BMJ* 2013
Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297-302.

Proportion of specific agreement

- Problem with normal % agreement

		Rheumatologist A		
		Progression	No progression	Total
Rheumatologist B	Progression	[a] 33	[b] 4	37
	No progression	[c] 4	[d] 5	9
	Total	37	9	46 (N)

Normal % agreement

$$Agreement_{\%} = \frac{a+d}{N} = \frac{33+5}{46} = 82{,}6\%$$

No progression
Number of misclassifications (b and c) outweigh the the numbers on which they do agree (d)

Progression
Number of misclassifications (b and c) is OK compared to the numbers where they agree (a)

This is NOT reflected in the normal % agreement but IS in proportion specific agreement

Proportion of specific agreement

		Rheumatologist A		
		Progression	No progression	Total
Rheumatologist B	Progression	[a] 33	[b] 4	37
	No progression	[c] 4	[d] 5	9
	Total	37	9	46 (N)

Observed prevalence of progression

$$Prevalence_{Prog} = \left(\left(\frac{a+c}{N}\right) + \left(\frac{a+b}{N}\right)\right) / 2 = \left(\left(\frac{33+4}{46}\right) + \left(\frac{33+4}{46}\right)\right) / 2 = 0.804$$

- Specific agreement on a positive rating (progression)

$$PA = \frac{2a}{(2a+b+c)} = \frac{2\times33}{((2\times33)+4+4)} = \frac{66}{74} = 0.892$$

➡ Likelyhood that another clinician would also rate progression

- Specific agreement on a negative rating (no progression)

$$NA = \frac{2d}{(2d+b+c)} = \frac{2\times5}{((2\times5)+4+4)} = \frac{10}{18} = 0.556$$

➡ Likelyhood that another clinician would also rate no progression

Measurement error

Scale	Parameters
Continuous	Standard Error of the Measurement (SEM) Bland & Altman - Limits of Agreement (LOA)
Ordinal	Proportion of specific agreement
Nominal	Proportion of specific agreement

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

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

47

Standard error of the measurement (SEM)

Definition

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

Purpose

- Quantify the extent 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

48

SEM

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

Coming later…

Example – cont'd

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

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

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

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

49

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

50

$SEM_{consistency}$

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

$$SEM_{consistency} = \sqrt{\sigma^2_{error}}$$

$$SEM_{consistency} = SD_{pooled} \times \sqrt{(1 - ICC_{consistency})}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

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

de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;59(10):1033–9.

$SEM_{agreement}$

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

$$SEM_{agreement} = \sqrt{\sigma^2_{measurements} + \sigma^2_{error}}$$

$$SEM_{agreement} = SD_{pooled} \times \sqrt{(1 - ICC_{agreement})}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;59(10):1033–9.

Bland & Altman LOA

Graphical presentation of mean and difference between two scores

- Limits of agreement (LOA) = measurement error

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

Systematic error

Random error

95% of the differences lies between the LOA – prediction interval

Three conditions

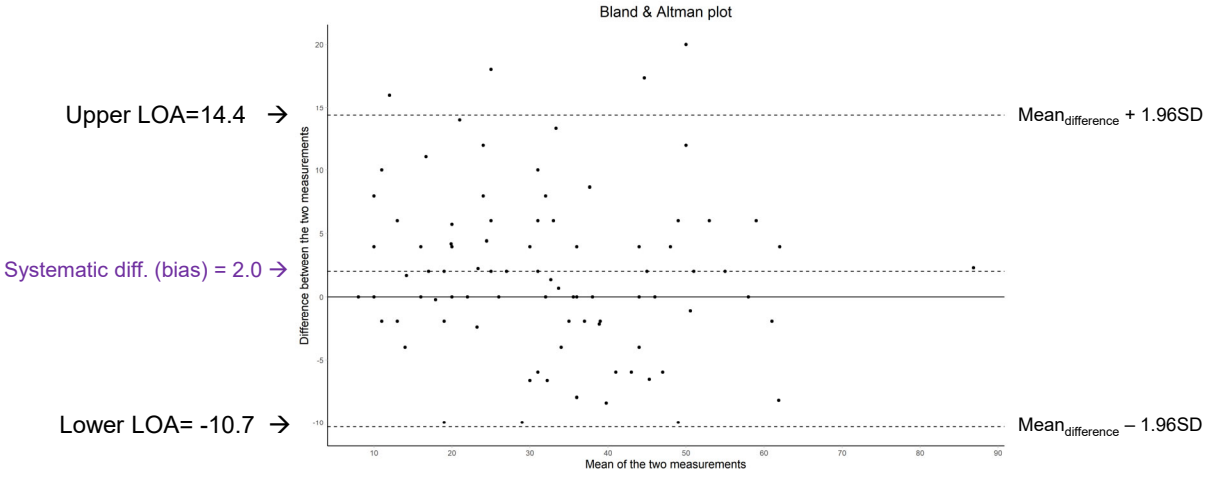
Assumptions of the Bland & Altman method:

- The differences should be independent
  - Checked by examining the study design
- The variability of the measurement errors is uniform throughout the whole range of the measurements
  - Checked by inspection of the Bland & Altman plot and/or normality plots of the differences
- The bias (systematic difference) between the measurement 1 and 2 is constant
  - Checked by regression method (Taffé method) splitting the bias into:
    - Differential bias (constant bias)
    - Proportional bias (non-constant bias) ⟹ Assumed to be zero in a test-retest situation

1. Bland JM, Altman DG. Measuring agreement in method comparison studies. StatMethods Med Res 1999;8(2):135–60.
2. Taffé P. When can the Bland & Altman limits of agreement method be used and when it should not be used. J Clin Epidemiol. 2021 Apr.

Agreement (test-retest design)

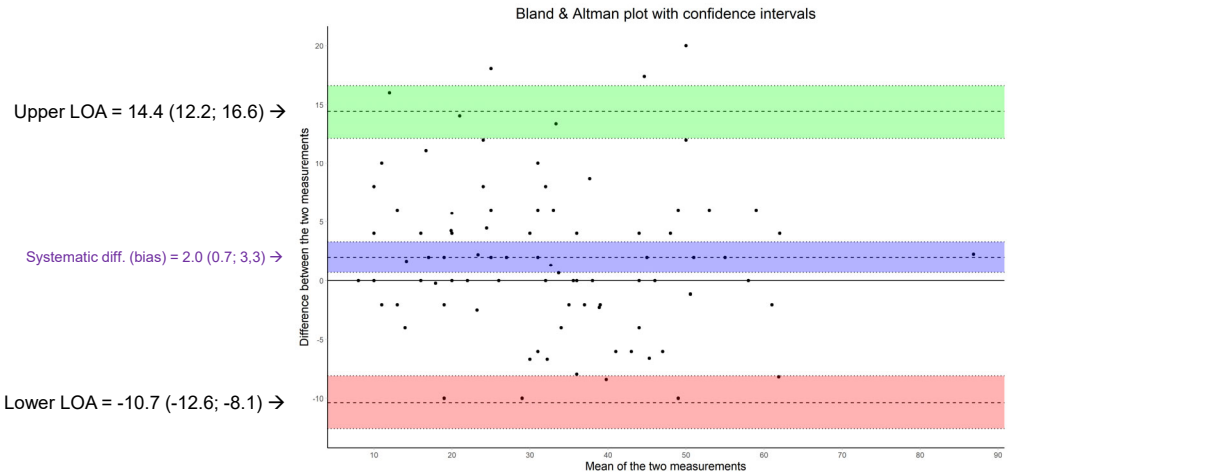
Oswestry Disability Index (n = 93)



55

Agreement (test-retest design)

Oswestry Disability Index (n = 93)



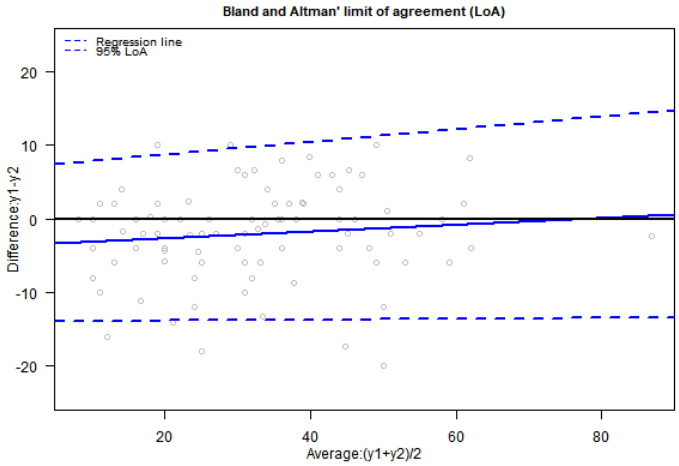
56

Agreement (test-retest design)

The 'Taffé method':

Used if assumptions are not met

E.g.,
Comparing the sensitivity of MRI and CT scans when detecting bone tumors



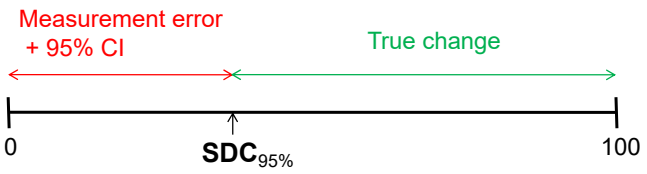
1. Taffé P. Assessing bias, precision, and agreement in method comparison studies. Stat Methods Med Res. 2020 Mar 1;29(3):778–96.
2. Taffé P, Halfon P, Halfon M. A new statistical methodology overcame the defects of the Bland–Altman method. J Clin Epidemiol. 2020 Aug 1;124:1–7.

57

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

de Vet et al. (2006) When to use agreement versus reliability measures. Journal of Clinical Epidemiology.

58

SEM, LOA and SDC

NB: $SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$

SEM

$$SEM_{consistency} = \sqrt{\sigma^2_{error}}$$

$$SEM_{consistency} = SD_{pooled}\sqrt{(1 - ICC_{consistency})}$$

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

$$SEM_{agreement} = \sqrt{\sigma^2_{measurements} + \sigma^2_{residual}}$$

$$SEM_{agreement} = SD_{pooled} \times \sqrt{(1 - ICC_{agreement})}$$

Limits of agreement (LOA)

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

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

Smallest Detectable Change ($SDC_{95\%}$)

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

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

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

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

59

SDC at group level

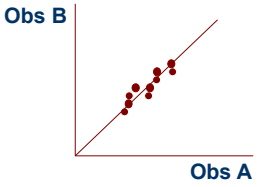
$$SDC_{agreement\ (individual)} = 1.96 \times SEM_{agreement} \times \sqrt{2}$$

$$SDC_{agreement\ (group)} = SDC_{agreement\ (individual)} \div \sqrt{n}$$

Franke TPC, de Vet HCW, Huisstede BMA. Minimally important change and smallest detectable change of the OSTRC questionnaire in half- and full-marathon runners. Scandinavian Journal of Medicine & Science in Sports. 2021;31(5):1048–58.

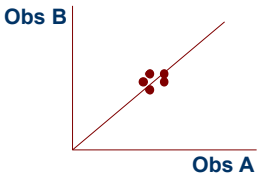
60

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



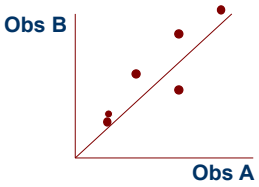
Rel = 0.85

SEM = 1.48



Rel = 0.50

SEM = 1.48



Rel = 0.98

SEM = 4.39

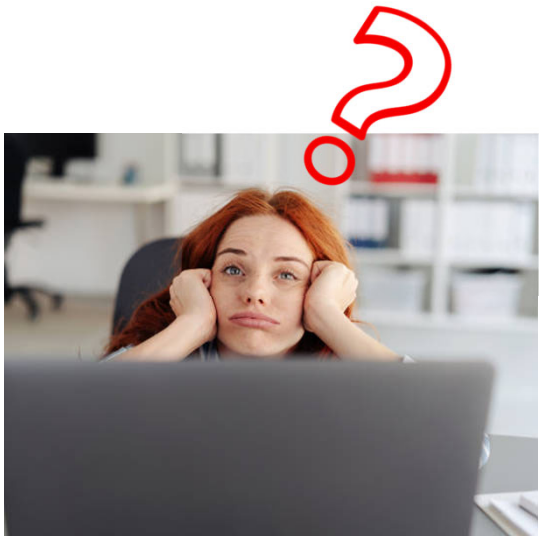
Stratford et al., 1989

61

Summary

Reproducibility

- Reliability
  - Measurement error related to the variation between patients
  - *Discriminative instruments*
  - Parameters: ICC, Kappa, PABAK, Gwet's AC1
- Measurement error
  - Measurement error on repeated measurements
  - *Evaluative instruments*
  - Parameters: Proportion specific agreement, SEM, LOA, SDC



62