NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

Measuring agreement between raters

by
Stian Lydersen

Presentation at NTNU
24 January 2014

Slides Updated 22 January 2014

Examples

– X-rays rated by radiologists
– Claims for compensation after alleged birth trauma judged by medical experts.
– Video recordings of parent – child interaction. Emotional attachment scored by psychologists.

Measures of agreement:

– Categorical data:
  – Cohen's kappa, alternatives and generalizations.
– Continuous data:
  – Intraclass correlation coefficient (ICC), different versions

Gisev et al (2013), Table 2:

Examples of interrater indices suitable for use with various types of data (not exhaustive)

	Level of measurement		
	Nominal / categorical	Ordinal	Interval and ratio
2 raters	Cohen's kappa ICC Weighted kappa	Weighted kappa ICC	Bland-Altman plots ICC
>2 raters	Fleiss' kappa ICC	Kendall's coefficient of concordance ICC	ICC

Categorical data:
Cohen's Kappa

Table 14.8 Assessments of 85 xeromammograms by two radiologists (from Boyd et al., 1982).

	Rater 2				
Rater 1	Normal	Benign	Suspected cancer	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected cancer	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

7

Table 14.7 The general counts of assessments by 2 raters using c categories.

	Rater 2				
Rater 1	1	2	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
c	n_{c1}	n_{c2}	$\cdots$	n_{cc}	n_{c+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	N

8

The general probabilities of assessments by 2 raters using c categories.

	Rater 2				
Rater 1	1	2	$\cdots$	c	Total
1	p_{11}	p_{12}	$\cdots$	p_{1c}	p_{1+}
2	p_{21}	p_{22}	$\cdots$	p_{2c}	p_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
c	p_{c1}	p_{c2}	$\cdots$	p_{cc}	p_{c+}
Total	p_{+1}	p_{+2}	$\cdots$	p_{+c}	1

9

Now, consider a situation where two raters each classify subjects in c categories, numbered from 1 to c. Let p_{ij} denote the probability that a subject is classified in catergories i and j by rater 1 and 2, respectively. An intuitive measure of agreement is the probability that the raters agree, which is

$$p_a = p_{11} + p_{22} + \ldots + p_{cc}. \qquad (0.1)$$

But part of this agreement is due to chance. Suppose that rater 1 assigns to category i with probability $p_{i+} = \sum_{j=1}^{c} p_{ij}$, and rater 2 assigns to category j with probability $p_{+j} = \sum_{i=1}^{c} p_{ij}$ independently of rater 1. Then, Cohen's probability of agreement by chance is given by

$$p_e = p_{1+}p_{+1} + p_{2+}p_{+2} + \ldots + p_{c+}p_{+c}. \qquad (0.2)$$

Cohen's kappa is defined as the relative proportion of agreements exceeding that by chance, which is

$$\boxed{\kappa = \frac{p_a - p_e}{1 - p_e}}. \qquad (0.3)$$

10

Example: Table 14.7:

Estimated agreement proportion:

$$\hat{p}_a = (21+17+15+1)/85 = 54/85 = 0.64$$

Cohen's probability of agreement by chance:

$$\hat{p}_e = (28\times33 + 38\times22 + 16\times29 + 3\times1)/85^2 = 0.31,$$

Cohen's kappa:

$$\hat{\kappa} = \frac{0.64 - 0.31}{1 - 0.31} = 0.47.$$

11

If only two categories:

Table 14.9 Assessments of 85 xeromammograms by two radiologists, dichotomized in two categories based on Table 14.8.

	Rater 2		
Rater 1	Normal or benign	Suspected cancer or cancer	Total
Normal or benign	54	1	55
Suspected cancer or cancer	12	18	30
Total	76	19	85

$\hat{\kappa} = 0.63$ for two categories

$\hat{\kappa} = 0.47$ when using all four categories.

A weighted kappa, described later, may be more appropriate for ordered categories.

12

Interpretation of kappa values

Table 14.10 Guidelines for interpreting kappa, Landis and Koch (1977).

Value of κ	Strength of agreement
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

Recommendation:
Show the original table data, not only the measure of agreement.

13

Confidence intervals for Cohen's kappa

The approximate standard error of kappa for dichotomous or nominal categories is given by Altman et al. (2000) as

$$\widehat{SE}(\hat{\kappa}) = \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{N(1-\hat{p}_e)^2}}, \qquad (0.1)$$

An approximate $1-\alpha$ confidence interval is given by $\hat{\kappa} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\kappa})$.

A 95% CI based on the data in Table 14.9 is (0.45, 0.82). Some software uses other formulae, see Lydersen (2012) and references therein.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

14

Cohen's kappa:
Unexpected results or paradoxes.

- Depends on the number of categories, especially for nominal categories
- Depends on the marginal distribution (prevalence) of the categories
- Raters who disagree more on the marginal distribution may produce higher kappa values

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

15

Kappa depends on the marginal distribution:

Table 14.9

	Rater 2		
Rater 1	Normal	Cancer	Total
Normal	54	1	55
Cancer	12	18	30
Total	76	19	85

$\hat{\kappa} = 0.63$

Table 14.10

	Rater 2		
Rater 1	Normal	Cancer	Total
Normal	68	1	69
Cancer	12	4	16
Total	80	5	85

$\hat{\kappa} = 0.32$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

16

Raters who disagree more on the marginal distribution may produce higher kappa values:

Table 14.11: Symmetrical imbalance

	Rater 2		
Rater 1	disease	healthy	Total
disease	50	10	60
healthy	20	20	40
Total	70	30	100

$\hat{\kappa} = 0.35$

Table 14.12: Asymmetrical imbalance
(Raters disagree on which state is most prevalent)

	Rater 2		
Rater 1	disease	healthy	Total
disease	30	30	60
healthy	0	40	40
Total	30	70	100

$\hat{\kappa} = 0.44$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

17

Cohens weighted kappa:

Weights the degree of agreement (distance from the diagonal)

Linear weighted kappa: $w_{ij} = 1 - \frac{|i-j|}{c-1}$

With 4 categories, the weights are 1 on the diagonal, and 2/3, 1/3 and 0 off the diagonal.

Quadratic weighted kappa: $w_{ij} = 1 - \frac{(i-j)^2}{(c-1)^2}$

With 4 categories, the weights are 1 on the diagonal, and 8/9, 5/9 and 0 off the diagonal.

Unweighted kappa:

The weights are 1 on the diagonal, and always 0 off the diagonal

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

18

Unweighted

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Quadratic

1	8/9	5/9	0
8/9	1	8/9	5/9
5/9	8/9	1	8/9
0	5/9	8/9	1

Linear

1	2/3	1/3	0
2/3	1	2/3	1/3
1/3	2/3	1	2/3
0	1/3	2/3	1

User-defined (example)

1	.8	0	0
.8	1	0	0
0	0	1	.8
0	0	.8	1

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

19

Linear versus quadratic weighted kappa?

- No clear advice in the literature
- For the case of equal marginal distributions, that is, $n_{i+} = n_{+i}$ for all i, then the quadratic weighted $\hat{\kappa}_w$ is equal to the *intraclass correlation coefficient* $\widehat{ICC}_2$ described in Section 14.8, except for a term involving the factor $1/N$

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

20

Table 14.8 Assessments of 85 xeromammograms by two radiologists

	Rater 2				
Rater 1	Normal	Benign	Suspect cancer	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected cancer	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

Unweighted kappa: 0.47
Linear weighted kappa: 0.57
Quadratic weighted kappa: 0.67
User-defined (example) 0.59

Dichotomized table kappa (Table 14.9): 0.63

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

21

Categorical data: Alternatives to Cohen's kappa

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

22

Alternative measures, two raters:

- Assuming independence between raters:
  - Cohen's kappa (1960)
  - Scott's pi (1955)
  - Bennet's sigma (1954)
- Assuming some subjects are easy, other difficult to agree on:
  - Gwets AC1 (Gwet's gamma) (2001, 2008)
  - Aickin's alpha (1990)
  - Martin and Femia's Delta (2004, 2008) for multiple choice tests

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

23

Measures which differ only in terms of calculating chance agreement:

Cohen's kappa (1960) uses the product of the marginals,

$$\hat{p}_e = \sum_{i=1}^{c} \hat{p}_{i+}\hat{p}_{+i}$$

where $\hat{p}_{i+} = n_{i+}/n$, and $\hat{p}_{+i} = n_{+i}/n$

Scott's pi (1955) uses the squared average of the marginals,

$$\hat{p}_e = \sum_{i=1}^{c} \left[\left(\hat{p}_{i+} + \hat{p}_{+i} \right)/2 \right]^2$$

Bennet's sigma (1954) assumes a uniform marginal:

$$\hat{p}_e = 1/c$$

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

24

Gwet's gamma (2001, 2008) (Also called Gwet's AC_1):

$$\hat{p}_e = \frac{1}{c-1} \sum_{i=1}^{c} \hat{p}_i \left(1 - \hat{p}_i\right),$$

where $\hat{p}_i = \left(\hat{p}_{i+} + \hat{p}_{+i} \right)/2$, $\hat{p}_{i+} = n_{i+}/n$, and $\hat{p}_{+i} = n_{+i}/n$

When $c = 2$, the equation reduces to

$$\hat{p}_e = 2\hat{p}_1\hat{p}_2.$$

NTNU Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

25

Gwet's gamma and Aickin's alpha:

Easy subjects to classify (E) will be classified (deterministic) in the same category by both raters.

Hard subjects to classify (H) will be random classified. Probability $1/c$ for each of the c categories.

Aickin assumes each subject is either hard for both raters (HH), or easy for both raters (EE).

Gwet allows also a subject to be hard for Rater 1 and easy for Rater 2 (HE), or vice versa (EH)

26

Possible outcomes with Gwet's theory (Gwet, 2012):

Table 4.3:
Distribution of Population Subjects by Sub-Population of H- and E-Subjects, by Rater, and by Response Category (1,2)

Rater A		Rater *B*: Hard Subjects 1	Hard Subjects 2	Easy Subjects 1	Easy Subjects 2	Total	
Hard Subjects	1	N_{11}^{HH}	N_{12}^{HH}	N_{11}^{HE}	N_{12}^{HE}	N_{1+}^{H}	N_{HH+}
	2	N_{21}^{HH}	N_{22}^{HH}	N_{21}^{HE}	N_{22}^{HE}	N_{2+}^{HH}	
Easy Subjects	1	N_{11}^{EH}	N_{12}^{EH}	N_{11}^{EE}	0	N_{1+}^{E}	N_{E+}
	2	N_{21}^{EH}	N_{22}^{EH}	0	N_{22}^{EE}	N_{2+}^{E}	
Total		N_{+1}^{H}	N_{+2}^{H}	N_{+1}^{E}	N_{+2}^{E}	N	
		N_{+H}		N_{+E}			

27

Possible outcomes with Aickin's theory (Gwet, 2012):

Table 4.2:
Distribution of N Population Subjects by Rater, Subpopulation, and Response Category.

Rater A		Rater *B*: Hard Subjects 1	Hard Subjects 2	Easy Subjects 1	Easy Subjects 2	Total	
Hard Subjects	1	$N_{11}^{(H)}$	$N_{12}^{(H)}$			$N_{1+}^{(H)}$	N_H
	2	$N_{21}^{(H)}$	$N_{22}^{(H)}$			$N_{2+}^{(H)}$	
Easy Subjects	1			$N_{1}^{(E)}$	0	$N_{1}^{(E)}$	N_E
	2			0	$N_{2}^{(E)}$	$N_{2}^{(E)}$	
Total		$N_{+1}^{(H)}$	$N_{+2}^{(H)}$	$N_{1}^{(E)}$	$N_{2}^{(E)}$	N	
		N_H		N_E			

28

The inter-rater reliability measures (to be estimated) can be expressed as below. These expressions are definitional, since N_{ii}^{EE} etc are not observed.

Gwet's gamma:

$$\gamma_1 = \frac{\sum_{i=1}^{c} N_{ii}^{EE}}{N - \left(\sum_{i=1}^{c} N_{ii}^{HH} + \sum_{i=1}^{c} N_{ii}^{HE} + \sum_{i=1}^{c} N_{ii}^{EH}\right)}$$

Aickin's alpha:

$$\alpha = \frac{\sum_{i=1}^{c} N_{ii}^{EE}}{N}$$

29

Gwet's γ_1:
Green framed in numerator. All except crossed out in denominator.

Table 4.3:
Distribution of Population Subjects by Sub-Population of H- and E-Subjects, by Rater, and by Response Category (1,2)

Rater A		Rater *B*: Hard Subjects 1	Hard Subjects 2	Easy Subjects 1	Easy Subjects 2	Total	
Hard Subjects	1	$\cancel{N_{11}^{HH}}$	N_{12}^{HH}	$\cancel{N_{11}^{HE}}$	N_{12}^{HE}	N_{1+}^{H}	N_{HH+}
	2	N_{21}^{HH}	$\cancel{N_{22}^{HH}}$	N_{21}^{HE}	$\cancel{N_{22}^{HE}}$	N_{2+}^{HH}	
Easy Subjects	1	$\cancel{N_{11}^{EH}}$	N_{12}^{EH}	$\boxed{N_{11}^{EE}}$	0	N_{1+}^{E}	N_{E+}
	2	N_{21}^{EH}	$\cancel{N_{22}^{EH}}$	0	$\boxed{N_{22}^{EE}}$	N_{2+}^{E}	
Total		N_{+1}^{H}	N_{+2}^{H}	N_{+1}^{E}	N_{+2}^{E}	N	
		N_{+H}		N_{+E}			

30

Aickin's α:
Green framed in numerator. All in denominator.

Table 4.2:
Distribution of N Population Subjects by Rater, Subpopulation, and Response Category.

Rater A		Rater *B*: Hard Subjects 1	Hard Subjects 2	Easy Subjects 1	Easy Subjects 2	Total	
Hard Subjects	1	$N_{11}^{(H)}$	$N_{12}^{(H)}$			$N_{1+}^{(H)}$	N_H
	2	$N_{21}^{(H)}$	$N_{22}^{(H)}$			$N_{2+}^{(H)}$	
Easy Subjects	1			$\boxed{N_{1}^{(E)}}$	0	$N_{1}^{(E)}$	N_E
	2			0	$\boxed{N_{2}^{(E)}}$	$N_{2}^{(E)}$	
Total		$N_{+1}^{(H)}$	$N_{+2}^{(H)}$	$N_{1}^{(E)}$	$N_{2}^{(E)}$	N	
		N_H		N_E			

31

Multiple choice tests:
Assume the student knows, say, 40% of the answers ($\Delta = 0.4$). He/she will answer 40% correct, and randomly choose the answers for the remaining questions.

Martin and Femia (2004) suggested this estimator:

$$\hat{\Delta} = \hat{p}_{11} + \hat{p}_{22} - 2\sqrt{\hat{p}_{12}\hat{p}_{21}}$$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

32

Table 14.9

Rater 1	Rater 2 Normal	Cancer	Total
Normal	54	1	55
Cancer	12	18	30
Total	76	19	85

$\hat{\kappa} = 0.635$, $\hat{\pi} = 0.627$, $\hat{\sigma} = 0.694$, $\gamma_1 = 0.741$, $\Delta = 0.766$

Table 14.10

Rater 1	Rater 2 Normal	Cancer	Total
Normal	68	1	69
Cancer	12	4	16
Total	80	5	85

$\hat{\kappa} = 0.320$, $\hat{\pi} = 0.294$, $\hat{\sigma} = 0.694$, $\gamma_1 = 0.805$, $\Delta = 0.766$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

33

Table 14.11: Symmetrical imbalance

Rater 1	Rater 2 disease	healthy	Total
disease	50	10	60
healthy	20	20	40
Total	70	30	100

$\hat{\kappa} = 0.348$, $\hat{\pi} = 0.341$, $\hat{\sigma} = 0.400$, $\gamma_1 = 0.450$, $\Delta = 0.417$

Table 14.12: Asymmetrical imbalance (Raters disagree on which state is most prevalent)

Rater 1	Rater 2 disease	healthy	Total
disease	30	30	60
healthy	0	40	40
Total	30	70	100

$\hat{\kappa} = 0.444$, $\hat{\pi} = 0.394$, $\hat{\sigma} = 0.400$, $\gamma_1 = 0.406$, $\Delta = 0.700$ (or 0.585)

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

34

Gwet's gamma is paradox-resitant (Gwet, 2012)

Wongpakaran, Wongpakaran, Wedding and Gwet (2013):
"It is interesting to note that although Gwet proved that the AC1 is better than Cohen's Kappa in 2001, a finding subsequently confirmed by biostatisticians [18], few researchers have used AC1 as a statistical tool, or are even aware of it, especially in the medical field. "

But ref [18] only illustrates that AC1 is resistant to the prevalence paradox.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

35

Comparisons of measures for 2 raters:

(Ato, Lopez, & Benavente 2011) compare measures in terms of their ability to estimate the systematic agreement proportion.
Hence, the construct (estimand) is Δ (?).
Recommend Bennet's sigma, and Martin and Femia' Delta (of course), since these have least bias.

(Wongpakaran, Wongpakaran, Wedding, & Gwet 2013) compare Cohen's kappa and Gwet's gamma.
"Our results favored Gwet's method over Cohen's kappa with regard to prevalence or marginal probability problem."

BUT:
- The different measures estimate different constructs!
- In reality, subjects are somewhere on a continuous scale from easy to completely random to rate.

SO:
It is not obvious which measure is "best"!

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

36

Categorical data: Generalizations to more than two raters

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

37

More than two raters

- No unique way to generalize
- Fleiss' kappa (1971) is a generalization of Scott's pi
- Conger's chance agreement probability (1980) is a generalization of Cohen's kappa. Computations are time-consuming if more than three raters.
- Gwet (2012, page 31) recommends using Fleiss' kappa

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

38

Continuous data: Intraclass correlation coefficient *(ICC)*

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

39

The intraclass correlation coefficient *(ICC)*

- Measures the correlation between one measurement on a subject and another measurement on the same subject (Shrout and Fleiss, 1979).
- Several *ICC* versions exist for different study designs and study aims
- The term *ICC* is also used in other settings, such as replicated measurements per subject, or patients within clinics.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

40

Three study designs:

Case 1:
Each subject is rated by a different set of *k* raters, randomly selected from a larger population of raters.

Case 2:
A random sample of *k* raters is selected from a larger population of raters. Each subject is rated by each rater.

Case 3:
There are only *k* raters of interest. Each subject is rated by each rater.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

41

Table 14.13 Definitions of *ICC* with different models and notations used by different authors. The *ICC* measures with *k* in parentheses are defined for the average of *k* measurements, and the others are for single measurements.

ANOVA model	Interaction between rater and subject?	Authors: Shrout and Fleiss (1979)	Authors: McGraw and Wong (1996)	Authors: Barnhart et al. (2007)
One-way random effects		Case 1 $ICC(1,1)$ or $ICC(1,k)$	Case 1 $ICC(1)$ or $ICC(k)$	ICC_1
Two-way random effects	Without interaction	As below	Case 2A $ICC(A,1)$ or $ICC(A,k)$	ICC_2
	With interaction	Case 2 $ICC(2,1)$ or $ICC(2,k)$	Case 2 As above	ICC_3
Two-way mixed effects	Without interaction	As below	Case 3A $ICC(A,1)$ or $ICC(A,k)$	ICC_2
	With interaction	Case 3 $ICC(3,1)$ or $ICC(3,k)$	Case 3 As above	ICC_3

Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

42

We shall limit our focus to agreement between single measurements, without interaction, and we use the notation *ICC1* and *ICC2* of Barnhart et al. (2007) in Table 14.14.

Alternatively, agreement can be defined for average of *k* measurements.

The intraclass correlation *ICC(3,k)* in Table 14.14 is equivalent to Cronbach's alpha, a commonly used measure of the internal consistency of items on a psychometric scale.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

43

Case 1:

One-way random effect model

$X_{ij} = \mu + b_i + w_{ij}$

where

X_{ij} is rating number j on subject number i,

$b_i \sim N(0, \sigma_B^2)$ is the random effect of subject number i,

$w_{ij} \sim N(0, \sigma_W^2)$ is a residual term.

In case 1, 2, and 3, all random effects and residual terms are assumed independent.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

44

The correlation between two ratings X_{ij_1} and X_{ij_2} on subject number i is

$$ICC_1 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

45

In Case 1, w_{ij} includes a rater effect and an error term.

In cases 2 and 3, the components of w_{ij} are specified:

$$X_{ij} = \mu + b_i + c_j + e_{ij},$$

where

c_j is the effect of rater j

e_{ij} is the residual random error.

Case 2: $c_j \sim N(0, \sigma_C^2)$

Case 3, c_j is a fixed effect with constraint $\sum_{j=1}^{k} c_j = 0$.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

46

The correlation between two ratings X_{ij_1} and X_{ij_2} on subject number i is

$$ICC_2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_C^2 + \sigma_E^2}.$$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

47

Be aware of:

- ICC, like Pearson’s correlation coefficient, is highly influenced by the variability in the subjects. The larger variation between subjects, and ICC will be closer to one.
- ICC combines any systematic difference between the raters and the random measurement variation, in one measure.
- If the purpose is to compare two measurement methods rather than two raters, Bland and Altman (1986) recommend to not use a correlation coefficient. Rather, they recommend plotting the difference between to measurements as a function of their mean, commonly termed a Bland-Altman plot.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

48

Example: Video recordings of parent – child interaction.

- An RCT of Marte Meo versus treatment as usual
- Three time points: Baseline, 2 months, and 8 months
- Emotional attachment (EA) score based on video recording of parent – child interaction. Rating scored by a psycholigist or psychiatrist.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

49

Design of Interrater reliability (IRR) study

- 36 distinct individuals, 12 from each of 3 time points.
- Each was rated by 2 raters, from a pool of 4 raters.
- All 6 combinations of raters rated 2 individuals at each of the 3 time points.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

50

Design ... (continued)

- Three first-raters (A, B, C) at each time point.
- Four second-raters at each time point (A, B, C, D)
- At each time point 12 pairs of raters.

AD	AB
BD	BA
CD	BC
AD	CB
BD	AC
CD	CA

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

51

Linear model with crossed random effects of individual and rater

Score on individual *i* by rater *j*:

$$X_{ij} = \beta_0 + \beta_1 time_2 + \beta_2 time_3 + b_i + c_j + e_{ij}$$

Analyzed in Stata as described by Rabe-Hesketh & Skrondal (2012), page 437-441.

(Show results from Word document)

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

52

Effect of rating 2 versus rating 1 on same individual?

$$X_{ij} = \beta_0 + \beta_1 time_2 + \beta_2 time_3 + \beta_3 rating_2 + b_i + c_j + e_{ij}$$

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

53

References

Ato, M., Lopez, J.J., & Benavente, A. 2011. A simulation study of rater agreement measures with 2x2 contingency tables. *Psicologica*, 32, (2) 385-402

Barnhart, H.X., Haber, M.J., & Lin, L.I. 2007. An overview on assessing agreement with continuous measurements. *J.Biopharm.Stat.*, 17, (4) 529-569

Gisev, N., Bell, J.S., & Chen, T.F. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res.Social.Adm Pharm.*, 9, (3) 330-338

Gwet, K.L. 2012. *Handbook of Inter-rater Reliability*, 3 ed. Gaithersburg, Maryland, Advanced Analytics, LLC.

Lydersen, S. 2012, "Diagnostic tests, ROC curves, and measures of agreement," *In Medical statistics in clinical and epidemiological research*, M. B. Veierød, S. Lydersen, & P. Laake, eds., Oslo: Gyldendal Akademisk, pp. 462-492.

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu

54

Martin, A. & Femia, P. 2004. Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Phychology* (57) 1-19

Mcgraw, K.O. & Wong, S.P. 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, (1) 30-46

Rabe-Hesketh, S. & Skrondal, A. 2012. Multilevel and longitudinal modeling using Stata, 3rd ed. College Station, Tex, Stata Press Publication.

Shrout, P.E. & Fleiss, J.L. 1979. Intraclass correlations: uses in assessing rater reliability. Psychol.Bull., 86, (2) 420-428

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K.L. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. Bmc Medical Research Methodology, 13

NTNU
Regional Centre for Child and Youth Mental Health and Child Welfare

www.ntnu.edu/rkbu