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

Measuring agreement between raters

by
Stian Lydersen

Presentation at RBUP, Oslo, 9 March 2016

Slides updated 6 March 2016
http://folk.ntnu.no/slyderse/medstat/Interrater_fullpage_9March2016.pdf
http://folk.ntnu.no/slyderse/medstat/Interrater_6slides_9March2016.pdf

Seminar:
Measuring agreement between raters.
by
Stian Lydersen, Professor of Medical Statistics, RKBU Midt.

RBUP Oslo, 9 March 2016, kl 0830 – 1130.

In many situations, subjects are rated by experts who use some degree of judgment. Examples are X-rays rated by radiologists, or patient video recordings given a score by psychologists. Since there is some element of judgment, the agreement between the raters will not be perfect.

Common measures of agreement will be presented. This includes Cohen's kappa and alternative measures for categorical data, and different versions of the intraclass correlation coefficient (ICC) for continuous data.

The choice of measure of agreement depends on:
a) The research question at hand
b) Whether there are two raters, or more than two raters
c) Whether the ratings are dichotomous, nominal, ordinal, or continuous
Recommendations for different situations will be given.

The presentation will be based on Section 14.6 – 14.8 in Lydersen (2012), Gisev et al. (2013), and recent examples from my own research.

Presentasjonen vil bli holdt på norsk dersom alle tilhørerne snakker norsk.

Examples

- X-rays rated by radiologists
- Claims for compensation after alleged birth trauma judged by medical experts. (Andreasen et al. 2014)
- Video recordings of parent – child interaction. Emotional attachment scored by psychologists. (Høivik et al. 2015)
- Psychiatric diagnosis based on Kiddie-SADS, based recorded telephone interview in the CAP (Hel-BUP) follow up study in Trondheim
- Retts-p: Rapid emergency triage and treatment system for children arriving at a pediatric emergency department. Categories red, orange, yellow, green. (Henning et al. 2016)

Measures of agreement:

- Categorical data:
  - Cohen's kappa, alternatives and generalizations.
  - Positive and negative agreement
- 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	Cohen's weighted kappa	Bland-Altman plots
	ICC	ICC	ICC
>2 raters	Fleiss' kappa	Kendall's coefficient of concordance	
	ICC	ICC	ICC

Categorical data: Cohen's Kappa

7

Table 14.7 Assessments of 85 xeromammograms by two radiologists (from Boyd et al., 1982).

Rater 1	Rater 2 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

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

www.ntnu.edu/rkbu

8

Table 14.6 The general counts of assessments by 2 raters using c categories.

Rater 1	Rater 2 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

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

www.ntnu.edu/rkbu

9

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

Rater 1	Rater 2 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

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

www.ntnu.edu/rkbu

10

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

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

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

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

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

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

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

www.ntnu.edu/rkbu

11

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:

$$\hat{p}_e = (28\times 33 + 38\times 22 + 16\times 29 + 3\times 1)/85^2 = 0.31,$$

Cohen's kappa:

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

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

www.ntnu.edu/rkbu

12

If only two categories:

Table 14.10 Assessments of 85 xeromammograms by two radiologists, dichotomized in two categories based on Table 14.7.

Rater 1	Rater 2 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.

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

www.ntnu.edu/rkbu

13

Interpretation of kappa values

Table 14.9 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.

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

www.ntnu.edu/rkbu

14

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

15

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

16

Kappa depends on the marginal distribution:.
Inter-rater reliability assessment, the CAP (Hel-BUP) study. (Schei et al. 2015)
28 participants (drawn randomly) were scored by two raters.

Anxiety — Cohen's kappa=0.50

		Rater 2		
		No	Yes	Total
Rater 1	No	19	2	21
	Yes	3	4	6
	Total	22	6	28

Psychotic — Cohen's kappa=0.0

		Rater 2		
		No	Yes	Total
Rater 1	No	27	1	28
	Yes	0	0	0
	Total	27	1	28

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

www.ntnu.edu/rkbu

17

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

18

Table 14.7 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

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

www.ntnu.edu/rkbu

19

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

20

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

21

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

22

Table 14.8 Assessments of 85 xeromammograms by two radiologists

Rater 1	Rater 2 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

23

Categorical data: Alternatives to Cohen's kappa

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

www.ntnu.edu/rkbu

24

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

25

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

26

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

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

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

27

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)

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

www.ntnu.edu/rkbu

28

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}			

Mental Health and Child Welfare

www.ntnu.edu/rkbu

29

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			

www.ntnu.edu/rkbu

30

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:

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

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

www.ntnu.edu/rkbu

31

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	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}			

32

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	$N_{11}^{(H)}$	$N_{12}^{(H)}$	[redacted]	[redacted]	$N_{1+}^{(H)}$	N_H
	2	$N_{21}^{(H)}$	$N_{22}^{(H)}$			$N_{2+}^{(H)}$	
Easy Subjects	1	[redacted]	[redacted]	$N_1^{(E)}$	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			

33

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

34

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

35

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

36

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.

The mathematics behind Gwet's gamma is difficult to follow. No clear justification for the use of Euclidian distance in the definition.
http://www.agreestat.com/book3/errors.html

NTNU

37

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

38

Categorical data: Generalizations to more than two raters

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

www.ntnu.edu/rkbu

39

More than two raters

– No unique way to generalize Cohen's kappa
– 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 (2014, page 52) recommends using Fleiss' kappa before Conger's chance agreement.
– There exists a generalization of Gwet's gamma to more than two raters

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

www.ntnu.edu/rkbu

40

More than two raters: Dichotomous data

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

www.ntnu.edu/rkbu

41

Example:
Andreasen, S., Backe, B., Lydersen, S., Øvrebø, K., & Øian, P. 2014.
The consistency of experts' evaluation of obstetric claims for compensation. BJOG., 122, (7) 948-953

The aim of this study was to investigate the consistency of experts' evaluation of different types of birth trauma, concerning malpractice, and causality between injury and the healthcare provided. Malpractice and causality qualifies for compensation.

In the questionnaire we presented 12 clinical scenarios concerning birth trauma to mother or child. All scenarios were based on real compensation claims to the NPE (Norsk Pasientskadeerstating).

In total, 14 medical experts participated.

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

www.ntnu.edu/rkbu

42

Software:

This free software turned out to have some errors:
http://www.statstodo.com/CohenKappa_Pgm.php

We used this commercial software:
www.agreestat.com

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

www.ntnu.edu/rkbu

Table 2. The experts' evaluation of negligence, causality, permanent injury and both negligence and causality in cases of injury during delivery. Estimates and 95% confidence intervals (CIs) for absolute agreement, Gwet's AC1 and Fleiss' kappa

Case	Negligence		Causality		Permanent injury		Negligence and causality**	
	Yes (n)	No (n)	Yes (n)	No (n)	Yes (n)	No (n)	Yes (n)	No (n)
Asphyxia								
3	13	1	13	1	*	*	12	2
5	14	0	13	1	13	1	13	1
6	10	4	4	10	4	10	4	10
8	13	1	4	9	4	9	4	10
11	10	4	1	13	1	13	1	13
Absolute agreement	0.77		0.73		0.70		0.71	
Gwet's AC1	0.69 (0.27–1.0)		0.47 (0.05–0.89)		0.43 (−0.09 to 0.95)		0.43 (0.07–0.79)	
Fleiss' kappa	0.05 (−0.06 to 0.16)		0.47 (0.05–0.88)		0.33 (−0.37 to 1.0)		0.43 (0.05–0.81)	
Sphincter tear (obstetric anal sphincter injury, OASIS)								
1	1	13	12	2	11	3	1	13
9	7	7	13	1	13	1	7	7
12	6	8	13	1	12	2	5	9
Absolute agreement	0.6		0.87		0.74		0.61	
Gwet's AC1	0.27 (−1.0 to 1.0)		0.78 (0.53–1.0)		0.65 (0.18–1.0)		0.32 (−1.0 to 1.0)	
Fleiss' kappa	0.09 (−0.41 to 0.59)		−0.06 (−0.08 to −0.05)		−0.05 (−0.12 to 0.02)		0.08 (−0.36 to 0.53)	
Hysterectomy								
4	4	10	5	9	6	8	4	10
10	13	0	13	0	13	0	13	0
Absolute agreement	0.78		0.75		0.74		0.78	
Gwet's AC1	0.59 (−1.0 to 1.0)		0.56 (−1.0 to 1.0)		0.55 (−1.0 to 1.0)		0.59 (−1.0 to 1.0)	
Fleiss' kappa	0.52 (−1.0 to 1.0)		0.43 (−1.0 to 1.0)		0.35 (−1.0 to 1.0)		0.52 (−1.0 to 1.0)	
Shoulder dystocia								
2	0	14	11	3	10	4	0	14
7	0	14	11	3	9	5	0	14
Absolute agreement	1.0***		0.64		0.53		1.0***	
Gwet's AC1	****		0.45 (0.45–0.45)*****		0.17 (−0.92 to 1.0)		****	
Fleiss' kappa	****		−0.08 (−0.08 to −0.08)*****		−0.07 (−0.08 to −0.07)		****	
Absolute agreement	0.77		0.74		0.69		0.74	
Gwet's AC1	0.54 (0.25–0.83)		0.54 (0.27–0.81)		0.47 (0.12–0.77)		0.52 (0.25–0.79)	
Fleiss' kappa	0.53 (0.24–0.82)		0.41 (0.20–0.61)		0.33 (0.10–0.57)		0.47 (0.17–0.75)	

*Numbers missing as question not relevant for the case.

44

	compensation		
case_n	0	1	Total
1	13	1	14
2	14	0	14
3	2	12	14
4	10	4	14
5	1	13	14
6	10	4	14
7	14	0	14
8	10	4	14
9	7	7	14
10	0	13	13
11	13	1	14
12	9	5	14
Total	103	64	167

45

	compensation		
expert	0	1	Total
1	8	4	12
2	5	7	12
3	8	4	12
4	6	6	12
5	8	3	11
6	6	6	12
7	7	5	12
8	9	3	12
9	7	5	12
10	7	5	12
11	7	5	12
12	8	4	12
13	7	5	12
14	10	2	12
Total	103	64	167

46

The probability to be judged eligible for compensation seems to:

Vary a lot between cases
Vary little between experts.

To quantify this, we used a logistic model with random effect of case_no and expert.

The random effects are crossed, not nested.

47

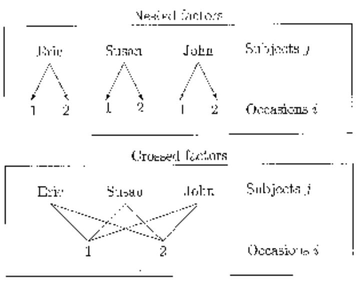


Figure 2.8: Illustration of nested and crossed factors

From Rabe- Hesketh and Skrondal, 2012, page 98

48

Logistic mixed model
(actually a two way random effects model):

$$p_{ij} = P(\text{Case no } i \text{ is classified as "1" by rater } j)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + b_i + c_j$$

where

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

$c_j \sim N(0, \sigma_C^2)$ is the random effect of rater j

49

Crossed random effects cannot (to my knowledge) be analyzed in SPSS. Possible in Stata as described in (Rabe-Hesketh & Skrondal 2012) page 437-441 and 900 – 907.

```
xtmelogit compensation || _all: R.case_no || expert :, var
estimates store expert_and_case
xtmelogit compensation || expert :, var
estimates store expert
lrtest expert expert_and_case
xtmelogit compensation || case_no:,var
estimates store case_no
lrtest case_no expert_and_case
```

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

www.ntnu.edu/rkbu

50

"A logistic model with the outcome that the experts stated malpractice and causality, gave the following results: The variance (on a log odds scale) between the 12 cases was 5.6 ($p<0.001$), and between the experts 0.009 ($p=1.0$). Hence, the probability to answer "yes" varies considerably between the cases, but practically does not vary between the experts."

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

www.ntnu.edu/rkbu

51

More than two raters: Ordinal data

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

www.ntnu.edu/rkbu

52

Ordinal measurement, more than two raters:

Henning, B., Lydersen, S., & Døllner, H. 2016. A reliability study of the rapid emergency triage and treatment system for children. *Scand.J.Trauma Resusc.Emerg.Med.*, 24, (1) 19

Retts-p: Rapid emergency triage and treatment system for children. Categories red, orange, yellow, green.

20 fictive cases, 19 nurses (wave 1), 12 nurses (wave 2, 12 months later)

Kendall's W is a rank correlation measure for k raters

If ρ is the average of Spearman's rho for all the k(k-1)/2 pairs of raters, then $W = \rho - (\rho-1)/k$ (Gwet, 2014, page 363)

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

www.ntnu.edu/rkbu

53

Table 2 Nurses' triage priority ratings of 20 fictive cases[a]

Fictive patient case number	Nurses triage priority ratings: Red N (%)	Orange N (%)	Yellow N (%)	Green N (%)	Total ratings N
1	1 (3)	29 (94)[b]	1 (3)	0	31
2	12 (40)	18 (60)[b]	0	0	30
3	0	0	3 (10)	27 (90)[b]	30
4	29 (97)[b]	1 (3)	0	0	30
5	10 (32)	19 (61)[b]	2 (6)	0	31
6	5 (17)	24 (83)[b]	0	0	29
7	2 (6)	29 (94)[b]	0	0	31
8	0	27 (90)[b]	3 (10)	0	30
9	30 (100)[b]	0	0	0	30
10	0	1 (3)	29 (94)[b]	1 (3)	31
11	1 (3)	26 (90)[b]	2 (7)	0	29
12	0	22 (76)[b]	7 (24)	0	29
13	0	11 (37)	19 (63)[b]	0	30
14	1 (3)	28 (97)[b]	0	0	29
15	0	2 (7)	28 (93)[b]	0	30
16	0	0	29 (94)[b]	2 (6)	31
17	0	6 (19)	6 (19)	19 (62)[b]	31
18	4 (13)	27 (87)[b]	0	0	31
19	0	5 (17)	24 (83)[b]	0	29
20	29 (97)[b]	1 (3)	0	0	30
Total	124 (21)	276 (46)	153 (25)	50 (8)	603

[a]In Study 1, Wave A (n = 367 ratings) and Wave B (n = 236 ratings) combined
[b]"Correct" priority rating as determined by the research group

Wave 1, 19 nurses:
W=0.822

Wave 2, 12 nurses:
W= 0.844

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

www.nt

54

Two raters, dichotomous data: *Positive and negative agreement*

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

www.ntnu.edu/rkbu

55

The paradox:

Cohen's kappa is low when most subjects are rated in one category (for example non-diseased) by both raters.

Possible solution:

Report two measures instead of one:
Positive agreement and negative agreement. (Cicchetti and Feinstein 1990;Feinstein and Cicchetti 1990). Cited 1443 times (per 1 March 2016)

Clinicians are right not to like Cohen's kappa. (de Vet et al. 2013)

Analogue to reporting sensitivity and specificity for diagnostic tests

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

www.ntnu.edu/rkbu

56

Sensitivity and specificity:
Compare test result and true disease status.

	Test result		
Disease status	Positive	Negative	Total
Diseased	a	b	$a+b$
Non-diseased	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

$\text{Sensitivity} = \frac{a}{a+b}, \quad \text{Specificity} = \frac{d}{c+d}$

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

www.ntnu.edu/rkbu

57

Example: Adolescents living in Recidential Youth Care institutions in Norway (Undheim et al., work in progress, 2016)

Affective disorders

CAPA (regarded as gold standard)	CBCL (by main contact) Positive	Negative	Total
Diseased	61	13	74
Non-diseased	69	70	139
Total	130	83	213

Sensitivity: 61/74=0.82 Specificity: 70/139=0.50

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

www.ntnu.edu/rkbu

58

Sensitivity and specificity: Can be computed if true disease status is known.
Positive and negative agreement when true status is unknown:

	Rater 2		
Rater 1	Positive	Negative	Total
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

$\text{Positive agreement} = \frac{a}{a+b/2+c/2}, \quad \text{Negative agreement} = \frac{d}{d+b/2+c/2}$

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

www.ntnu.edu/rkbu

59

Example revisited: The CAP (Hel-BUP) study

Anxiety

		Rater 2 No	Yes	Total
Rater 1	No	19	2	21
	Yes	3	4	6
	Total	22	6	28

Cohen's kappa=0.50

Positive agreement: 0.62

Negative agreement: 0.88

Psychotic

		Rater 2 No	Yes	Total
Rater 1	No	27	1	28
	Yes	0	0	0
	Total	27	1	28

Cohen's kappa=0.0

Positive agreement: 0.0

Negative agreement: 0.98

Schei et al. (2015)

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

www.ntnu.edu/rkbu

60

Continuous data: Intraclass correlation coefficient *(ICC)*

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

www.ntnu.edu/rkbu

61

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.

62

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.

63

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

64

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.

65

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.

66

The correlation between two ratings X_{ij_1} and X_{ij_2}

on subject number i is

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

67

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 $\sum_{j=1}^{k} c_j = 0$

68

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

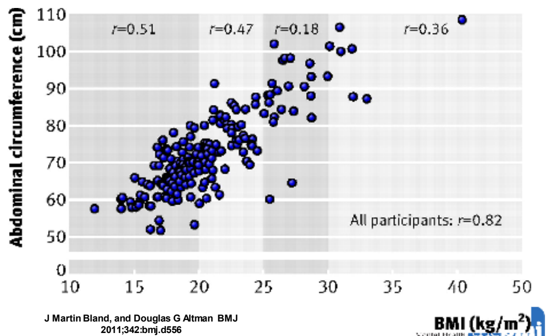
$$ICC_2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_C^2 + \sigma_E^2}.$$

69

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.

BMI and abdominal circumference in 202 men and women, with correlation coefficients in four restricted ranges and overall.



J Martin Bland, and Douglas G Altman BMJ 2011;342:bmj.d556

71

Example: Video recordings of parent – child interaction (Høivik et al., 2015)

- 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 psychologist or psychiatrist.

72

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.

73

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

74

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

75

There are 3 variance components:

Individual to be rated: 139.284=11.802^2
Rater: 22.973=4.793^2
Residual: 139.729=11.821^2

The total variance is
139.284 + 22.973 + 139.729 = 301.986 = 17.378^2

It follows (Rabe-Hesketh & Skrondal 2012, page 437-441) that the between rater, within individual intraclass correlation estimate is

$$\widehat{ICC} = \frac{139.284}{139.284 + 22.973 + 139.729} = 0.461$$

The average Pearson correlation between the raters was 0.63.

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

76

Effect of rating 2 versus rating 1 on same individual?

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

(Show results from word document)

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

77

The Cap Study (re-visited)

The IRR study was designed as follows: Seven of the interviewers were used as second opinion raters for taped telephone interviews. Each of these seven re-scored four interviews performed by four of the other six interviewers. Hence, the number of re-scored patients were 7x4=28. The design was constructed as shown in table 1, to be as balanced as possible.

			Second rater							
			A	B	C	D	E	F	G		Sum
First rater	A			1	1	1	1	0	0		4
	B		1		1	1	0	1	0		4
	C		0	0		1	1	1	1		4
	D		1	1	0		0	1	1		4
	E		1	0	0	1		1	1		4
	F		0	1	1	0	1		1		4
	G		1	1	1	0	1	0			4
	Sum		4	4	4	4	4	4	4		28

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

78

In the mixed effect model, the average CGAS score for rating number 1 was 74.07. For rating 2, the average score was 1.43 (p=0.31) higher. There are 3 variance components (given the fixed effect of rating number):

Individual to be rated: 187.0117 = 13.675^2
Rater: 9.789 = 3.129^2
Residual: 27.120 = 5.208^2

The total variance is
187.0117 + 9.789 + 27.120 = 223.9209 = 14.964^2

It follows (Rabe-Hesketh & Skrondal 2012, page 437-441) that the between rater, within individual intraclass correlation estimate is

$$\widehat{ICC} = \frac{187.0117}{187.0117 + 9.789 + 27.120} = 0.835$$

The variance between the raters was not statistically significant (Likelihood ratio test p=0.19). That is, there was no evidence that some raters tended to give systematically higher scores than others with respect to CGAS.

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

79

Henning et al (2016) revisited.

Retts-p: Rapid emergency triage and treatment system for children. Categories red (1), orange (2), yellow (3), green (4).

20 fictive cases, 19 nurses (wave 1), 12 nurses (wave 2, 12 months later)

$$ICC = \frac{Variance(patients)}{Variance(patients)+Variance(raters)+Residual}$$

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

www.ntnu.edu/rkbu

80

$$ICC = \frac{Variance(patients)}{Variance(patients)+Variance(raters)+Residual}$$

a linear mixed effect model including a fixed effect of Wave B, the estimated average rating at Wave A was 2.148, and the average rating at Wave B was 0.0439 (p = 0.168) higher. Since this is far from significant, we removed wave from the model. The average score of the reduced model was 2.208, and the total variance was $0.769 = 0.877^2$, including the variance between the rated patients ($0.627 = 0.792^2$), plus the variance due to the raters ($0.00212 = 0.046^2$) and the residual variance ($0.139 = 0.373^2$). The interrater reliability (ICC) estimate was 0.816.

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

www.ntnu.edu/rkbu

81

References

Andreasen, S., Backe, B., Lydersen, S., Øvrebo, K., & Øian, P. 2014. The consistency of experts' evaluation of obstetric claims for compensation. *BJOG*., 122, (7) 948-953

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

Cicchetti, D.V. & Feinstein, A.R. 1990. High agreement but low kappa: II. Resolving the paradoxes. *J.Clin.Epidemiol.*, 43, (6) 551-558

de Vet, H.C., Mokkink, L.B., Terwee, C.B., Hoekstra, O.S., & Knol, D.L. 2013. Clinicians are right not to like Cohen's kappa. *BMJ*, 346, f2125

Feinstein, A.R. & Cicchetti, D.V. 1990. High agreement but low kappa: I. The problems of two paradoxes. *J.Clin.Epidemiol.*, 43, (6) 543-549

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

www.ntnu.edu/rkbu

82

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. 2014. *Handbook of Inter-rater Reliability*, 4th ed. Gaithersburg, Maryland, Advanced Analytics, LLC.

Henning, B., Lydersen, S., & Døllner, H. 2016. A reliability study of the rapid emergency triage and treatment system for children. *Scand.J.Trauma Resusc.Emerg.Med.*, 24, (1) 19

Høivik, M.S., Lydersen, S., Drugli, M.B., Onsøien, R., Hansen, M.B., & Nielsen, T.S. 2015. Video feedback compared to treatment as usual in families with parent-child interactions problems: a randomized controlled trial. *Child Adolesc.Psychiatry Ment.Health*, 9, 3

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

83

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.

Schei, J., Nøvik, T.S., Thomsen, P.H., Lydersen, S., Indredavik, M.S., & Jozefiak, T. 2015. What Predicts a Good Adolescent to Adult Transition in ADHD? The Role of Self-Reported Resilience. *Journal of Attention Disorders*

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

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

www.ntnu.edu/rkbu

84

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