A Brief Tutorial on
Inter-Rater Agreement

Christian M. Meyer

Based on a tutorial on inter-rater agreement held as part of the doctoral program “Language and Knowledge Engineering” (LKE) at the Technische Universität Darmstadt, Germany on November 9, 2009 by Christian M. Meyer. All described measures have been implemented in DKPro Agreement.
https://code.google.com/p/dkpro-statistics/

Introduction
Validity, Reliability, Agreement

For each (manually or automatically generated) dataset, it is crucial to consider the following questions:

Is my evaluation valid?

- Can we draw conclusions from the data?
- One prerequisite for **validity** is that the evaluation data is **reliable**.

Is my evaluation data reliable?

- Is the generation reproducible?
- Raters annotate a sample of the data
- Assumption: The data is **reliable** if their **agreement** is good.

What is good agreement?

- How to measure agreement?
- How to interpret the result?
- **Inter-rater agreement coefficients**

Introduction
Notation

m raters $r \in R$
aka. coders, annotators, observers,…

matching?	*yes*	*yes*	*no*
score for..	*low*	*medium*	*low*
Apple	*NN*	*NNP*	*NN*
..bass..	*WN1*	*WN2*	*WN1*

n items $i \in I$ aka. units, records,…

>50 years of agreement studies – >50 different notation schemas!

k categories $c \in C$ aka. labels, annotations,…, which can be:
- binary (yes, no)
- ordinal (1, 2, 3,…)
- continuous (0.03, 0.49,…)
- ordered-category (low, high)
- nominal (NN, NNP, JJ, VB)
- likert-scale (strongly disagree, disagree, agree, strongly agree)

Percentage of Agreement
Definition

Contingency Matrix:

	r1 high	r1 low	
r2 high	2	2	**4**
r2 low	1	5	**6**
	3	**7**	**10**

Percentage of agreement:

$A_O = 1/n \, \Sigma_c \, (\#\ of\ agreements)$

$A_O = 1/10\ (2 + 5) = 0.7$

Relatedness?	r1	r2
gem – jewel	high	high
coast – shore	high	high
coast – hill	high	low
forest – graveyard	low	high
asylum – fruit	low	low
noon – string	low	low
automobile – wizard	low	low
brother – lad	low	high
cord – smile	low	low
autograph - shore	low	low

Example word pairs taken from Rubenstein & Goodenough (1965).
Calculation example is inspired by Artstein & Poesio (2008).

Percentage of Agreement
Standard Error and Confidence Interval

Contingency Matrix:

	r1 high	r1 low	
r2 high	2	2	**4**
r2 low	1	5	**6**
	3	**7**	**10**

Percentage of agreement:

$A_O = 1/n \, \Sigma_c \, (\# \text{ of agreements})$

$A_O = 1/10 \, (2 + 5) = 0.7$

Standard error:

$SE(A_O) = \sqrt{(A_O \, (1 - A_O))} \, / \, n$

$SE(A_O) = \sqrt{(0.7 \, (1 - 0.7))} \, / \, 10 = 0.04$

Confidence intervals:

$CL = A_O - SE(A_O) \cdot zCrit$

$CU = A_O + SE(A_O) \cdot zCrit$

$0.610 \leq 0.7 \leq 0.789$
with $zCrit = 1.96$ (95% confid.)

$0.624 \leq 0.7 \leq 0.775$
with $zCrit = 1.645$ (90% confid.)

Issues

Why is there Disagreement at all?

TECHNISCHE UNIVERSITÄT DARMSTADT

Sources of disagreement:

- **Insecurity** in deciding on a category
- **Hard/Debateable cases**
- **Carelessness**
- Difficulties or differences in **comprehending instructions**
- Openness for **distractions**
- Tendency to relax performance standard when **tired**
- personal **opinions/values**
- …

Possible corrective actions:

- **Training** of the annotators
- Write **better instructions**
- Provide better **environment**
- **Reduce amount** of annotated data per annotator
- Use **more annotators**
- …

Issues
Agreement by Chance

- Percentage of agreement does not regard **agreement by chance**
- Imagine the raters would guess randomly:

	r1 high	r1 low	
r2 high	45	45	**90**
r2 low	45	45	**90**
	90	**90**	**180**

$A_O = 1/180\ (45 + 45) = 0.5$

	r1 high	r1 med	r1 low	
r2 high	20	20	20	**60**
r2 med	20	20	20	**60**
r2 low	20	20	20	**60**
	60	**60**	**60**	**180**

$A_O = 1/180\ (20 + 20 + 20) = 1/3$

One would assume similar agreement
→ use chance-corrected measures!

Issues
Equal Weights

- **All categories are treated equally**
- Consider annotating/marking proper nouns in arbitrary texts

	r1 +	r1 –	
r2 +	10	20	**30**
r2 –	20	1,000	**1,020**
	30	**1,020**	**1,050**

$$A_O = 1/1050\ (10 + 1000) = 0.961$$

Almost perfect agreement, although the actual proper noun identification did not really work!

- For binary data: calculate **positive and negative agreement**

$$A_{O+} = 2\ (\#\ of\ agreements\ for\ +)\ /\ \Sigma_r\ (\#\ of + annotations)$$

$$A_{O+} = 2 \cdot 10\ /\ (30 + 30) = 0.333$$

$$A_{O-} = 2\ (\#\ of\ agreements\ for\ –)\ /\ \Sigma_r\ (\#\ of – annotations)$$

$$A_{O-} = 2 \cdot 1000\ /\ (1020 + 1020) = 0.980$$

(Cicchetti and Feinstein, 1990)

Issues
Summary

Measure	chance-corrected agreement	multiple raters	weighted categories
Percentage of Agreement	✘	✘	✘

Chance-corrected Measures
Definition

Basic idea: $$agreement = \frac{agreement\ beyond\ chance}{attainable\ chance\text{-}corrected\ agreement} = \frac{A_O - A_E}{1 - A_E}$$

Bennett, Alpert & Goldstein (1954)

$$S = \frac{A_O - A_E^{S}}{1 - A_E^{S}}$$

assume **uniform** distribution, i.e. the same probabilities for each categories:

$$A_E^{S} = \frac{1}{k}$$

Scott (1955)

$$\pi = \frac{A_O - A_E^{\pi}}{1 - A_E^{\pi}}$$

assume a **single distribution for all raters**, i.e. each rater annotates the same way:

$$A_E^{\pi} = \frac{1}{4n^2} \Sigma_c\, n_c^{2}$$

with the total number of annotations n_c for category c by all raters.

Cohen (1960)

$$\kappa = \frac{A_O - A_E^{\kappa}}{1 - A_E^{\kappa}}$$

assume **different probability distributions** for each rater:

$$A_E^{\kappa} = \frac{1}{n^2} \Sigma_c\, n_{c,r1}\, n_{c,r2}$$

with the total number of annotations $n_{c,r}$ for category c by rater r.

Chance-corrected Measures

Example

Basic idea: $agreement = \frac{A_O - A_E}{1 - A_E}$

	r1 high	r1 low	
r2 high	2	2	**4**
r2 low	1	5	**6**
	3	**7**	**10**

Percentage of agreement:

$A_O = 1/n\ \Sigma_c\ (\#\ of\ agreements)$

$A_O = 1/10\ (2 + 5) = 0.7$

Bennett et al.'s S:

$A_E^S = 1 / 2 = 0.5$

$S = (0.7 - 0.5) / (1 - 0.5) = 0.4$

Scott's π:

$A_E^\pi = 1/(4 \cdot 10^2)\ ((3 + 4)^2 + (6 + 7)^2)$
$= 0.545$

$\pi = (0.7 - 0.545) / (1 - 0.545) = 0.341$

Cohen's κ:

$A_E^\kappa = 1/10^2\ (3 \cdot 4 + 6 \cdot 7) = 0.54$

$\kappa = (0.7 - 0.54) / (1 - 0.54) = 0.348$

Issues
Agreement by Chance

- Percentage of agreement does not regard **agreement by chance**
- Imagine the raters would guess randomly:

	r1 high	r1 low	
r2 high	45	45	**90**
r2 low	45	45	**90**
	90	**90**	**180**

$A_O = 0.5$

	r1 high	r1 med	r1 low	
r2 high	20	20	20	**60**
r2 med	20	20	20	**60**
r2 low	20	20	20	**60**
	60	**60**	**60**	**180**

$A_O = 1/3$

$S = 0.0$
$\pi = 0.0$
$\kappa = 0.0$

Chance-corrected!

$S = 0.0$
$\pi = 0.0$
$\kappa = 0.0$

Issues
Summary

Measure	chance-corrected agreement	multiple raters	weighted categories
Percentage of Agreement	✘	✘	✘
Chance-corrected S	✓	✘	✘
Scott's π	✓	✘	✘
Cohen's κ	✓	✘	✘

Multiple Raters
Agreement Table

Relatedness?	r1	r2	r3
gem – jewel	high	high	high
coast – shore	high	high	low
coast – hill	high	low	high
forest – graveyard	low	high	high
asylum – fruit	low	low	high
noon – string	low	low	low
automobile – wizard	low	low	low
brother – lad	low	high	low
cord – smile	low	low	high
autograph - shore	low	low	high

convert to agreement table

Item	high	low
1	3	0
2	2	1
3	2	1
4	2	1
5	1	2
6	0	3
7	0	3
8	1	2
9	1	2
10	1	2

Example word pairs taken from Rubenstein & Goodenough (1965).

Multiple Raters
Generalized Measures

- So far, we only considered **two raters**, although there are usually more
- Generalize two-rater measures:

Fleiss (1971)

Generalizes Scott's π. The basic idea is to consider each pairwise agreement of raters and average over all items i.

$$\text{multi-}\pi = \frac{A'_O - A'^{\pi}_E}{1 - A'^{\pi}_E}$$

$$A'_O = \frac{1}{nm(m-1)} \Sigma_i \Sigma_c \, n_{i,c}(n_{i,c} - 1)$$

$$A'^{\pi}_E = \frac{1}{(nm)^2} \Sigma_c \, n_c^2$$

with the total number of raters $n_{i,c}$ that annotated item i with category c.

Davis and Fleiss (1982)

Generalizes Cohen's κ. The basic idea is to consider each pairwise agreement of raters and average over all items i.

$$\text{multi-}\kappa = \frac{A'_O - A'^{\kappa}_E}{1 - A'^{\kappa}_E}$$

$$A'_O = \frac{1}{nm(m-1)} \Sigma_i \Sigma_c \, n_{i,c}(n_{i,c} - 1)$$

$$A'^{\kappa}_E = \Sigma_c \frac{1}{\binom{m}{2}} \Sigma_{r1=1}^{m-1} \Sigma_{r2=r1+1}^{m} \frac{n_{r1,c} \, n_{r2,c}}{n^2}$$

with the total number of annotations $n_{c,r}$ by rater r for category c.

Multiple Raters

Example for *multi-π*

Fleiss (1971)

$$A'_O = \frac{1}{nm(m-1)} \Sigma_i \Sigma_c n_{i,c}(n_{i,c}-1)$$

$$A'_O = \frac{1}{10 \cdot 3(3-1)} 32 = 0.533$$

$$A'^{\pi}_E = \frac{1}{(nm)^2} \Sigma_c n_c^2$$

$$A'^{\pi}_E = \frac{1}{(10 \cdot 3)^2} (13^2 + 17^2) = 0.508$$

$$multi\text{-}\pi = \frac{A'_O - A'^{\pi}_E}{1 - A'^{\pi}_E} = 0.049$$

multi-π is also known as κ (Fleiss, 1971) and K (Carletta, 1996) – check definition!

$3 \cdot 2 + 0 \cdot (-1)$

$\Sigma_c n_{i,c}(n_{i,c}-1)$	Item	high	low
6	1	3	0
2	2	2	1
2	3	2	1
2	4	2	1
2	5	1	2
6	6	0	3
6	7	0	3
2	8	1	2
2	9	1	2
2	10	1	2
32	n_c	**13**	**17**

Issues
Summary

Measure	chance-corrected agreement	multiple raters	weighted categories
Percentage of Agreement	✘	✘	✘
Chance-corrected S	✔	✘	✘
Scott's π	✔	✘	✘
Cohen's κ	✔	✘	✘
multi-π	✔	✔	✘
multi-κ	✔	✔	✘

Krippendorff's α
Definition

- Allow further flexibility by allowing arbitrary **category metrics.**

Krippendorff (1980)

Derived from empirically statistics and content analysis. But can be represented in the same notation.

$$\alpha = 1 - \frac{D_O^\alpha}{D_E^\alpha} = \frac{\text{'est. var. within items'}}{\text{'est. total variance'}}$$

$$D_O^\alpha = \frac{1}{nm(m-1)} \Sigma_i \Sigma_{c1} \Sigma_{c2}\, n_{i,c1}\, n_{i,c2}\, d_{c1,c2}$$

$$D_E^\alpha = \frac{1}{nm(nm-1)} \Sigma_{c1} \Sigma_{c2}\, n_{c1}\, n_{c2}\, d_{c1,c2}$$

with the total number of raters $n_{i,c}$ that annotated item i with category c and the total number of annotations n_c for category c by all raters.

Distance function $d_{c1,c2}$

Arbitrary metric to allow working with

binary or nominal data:

$d_{c1,c2} = (c1 == c2\ ?\ 0\ :\ 1)$

with this distance function: $\alpha \approx \pi$

ordinal data ('square distance func.'):

$d_{c1,c2} = (c1 - c2)^2$

weighted data:

$d_{c1,c2}$	NN	NNP	VB
NN	–	0.1	0.9
NNP	0.1	–	0.9
VB	0.9	0.9	–

as well as interval, ratio data.

Krippendorff's α
Example

$n_{c1,c2}$	r1 +	r1 ●	r1 –	
r2 +	46	0	6	**52**
r2 ●	0	10	6	**16**
r2 –	0	0	32	**32**
	46	**10**	**44**	**100**

$d_{c1,c2}$	r1 +	r1 ●	r1 –
r2 +	0.0	0.5	1.0
r2 ●	0.5	0.0	0.5
r2 –	1.0	0.5	0.0

c	n_c
+	98
●	26
–	76

$n_{c1}\, n_{c2}\, d_{c1,c2}$	r1 +	r1 ●	r1 –
r2 +	0	1274	7448
r2 ●	1274	0	988
r2 –	7448	988	0

Krippendorff (1980)

$$D_O^\alpha = \frac{1}{nm(m-1)} \Sigma_i \Sigma_{c1} \Sigma_{c2}\, n_{i,c1}\, n_{i,c2}\, d_{c1,c2}$$

$$D_O^\alpha = \frac{46 \cdot 0 + 10 \cdot 0 + 6 \cdot 1 + 6 \cdot 0.5 + 32 \cdot 0}{100 \cdot 2(2-1)}$$

$$= 0.09$$

$$D_E^\alpha = \frac{1}{nm(nm-1)} \Sigma_{c1} \Sigma_{c2}\, n_{c1}\, n_{c2}\, d_{c1,c2}$$

$$D_E^\alpha = \frac{1274 + 7448 + 1274 + 988 + 7448 + 988}{100 \cdot 2(100 \cdot 2 - 1)}$$

$$= 0.4879$$

$$\alpha = 1 - \frac{D_O^\alpha}{D_E^\alpha} = 1 - \frac{0.09}{0.4879} = 0.8155$$

Issues
Summary

Measure	chance-corrected agreement	multiple raters	weighted categories
Percentage of Agreement	✘	✘	✘
Chance-corrected S	✔	✘	✘
Scott's π	✔	✘	✘
Cohen's κ	✔	✘	✘
multi-π	✔	✔	✘
multi-κ	✔	✔	✘
Krippendorff's α	✔	✔	✔
Weighted κ *(not covered here)*	✔	✔	✔

Side Note: Criticism

"The Myth of Chance-Corrected Agreement"

TECHNISCHE UNIVERSITÄT DARMSTADT

- Chance-corrected measures have also been criticized
- The presented measures S, π, κ assume that the raters are **completely statistically independent**
  - This means (1) the raters **guess on every item** or (2) the raters guess with probabilities **similar to the observed ratings**.
  - (1) is clearly not valid for an annotation study
  - (2) would not need a chance-correction
- Another argument is the different approach when **comparing to a gold standard** → measure precision/recall without any chance-correction
- John Uebersax proposes using raw agreements and focus on statistic significance tests, standard error and confidence intervals
- cf. (Uebersax, 1987; Agresti, 1992; Uebersax, 1993)

Traditional Statistics

Why not use χ^2 or correlations?

	r1 +	r1 ●	r1 –	
r2 +	25	13	12	**50**
r2 ●	12	2	16	**30**
r2 –	3	15	2	**20**
	40	**30**	**30**	**100**

Adapted from Cohen (1960)

$\chi^2 = 64.59$

$A_O = 0.36$
$S = 0.04$
$\pi = 0.02$
$\kappa = 0.04$

χ^2 is highly significant, because of the strong **associations**

+/+, ●/–, –/ ●

The **agreement** is however low!

A	B
1	1
2	2
3	3
4	4
5	5

Pearson correlation r vs. Cohen's κ:

$r = 1.0$
$\kappa = 1.0$

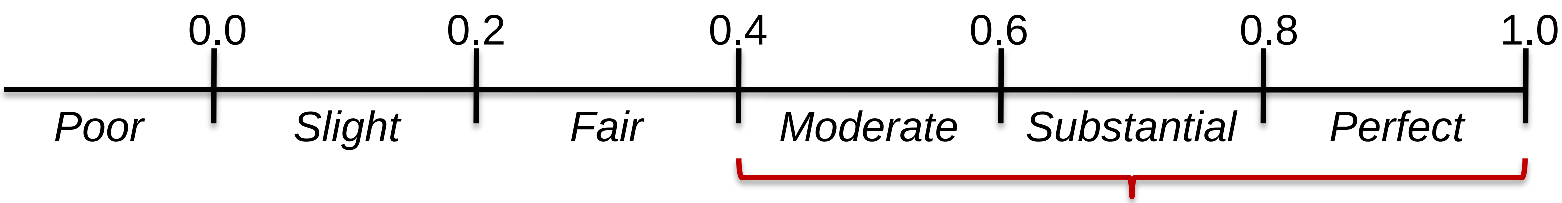
$r = 1.0$
$\kappa = -0.08$

Correlation measures are not suitable to measure inter-rater agreement!

A	B
1	2
2	4
3	6
4	8
5	10

Interpretation
What is good agreement?

- **Landis and Koch (1977)**



- **Krippendorff (1980), Carletta (1996)**
  - $0.67 < K < 0.8$ “allowing tentative conclusions to be drawn”
  - above 0.8 “good reliability”

- **Krippendorff (2004)**
  - “even a cutoff point of 0.8 is a pretty low standard”

- **Neuendorf (2002)**
  - “reliability coefficients of 0.9 or greater would be acceptable to all, 0.8 or greater […] in most situations”

Recommendations
by Artstein and Poesio (2008)

1. Anything is better than nothing
2. Give **details** on your study (who annotates and how?)
3. Use intensive **training** or professionals annotators
4. Report also the **agreement table/contingency matrix** rather than only the obtained agreement
5. Annotate with as **many raters** as possible, since it reduces the difference between the measures
6. Use $\boldsymbol{K}$ (equal to ***multi-π***) or $\boldsymbol{\alpha}$ which are used in the majority of studies, allow comparison and solve chance-related issues
7. Use Krippendorff's α for category labels that are not distinct from each other (**custom distance function**)
8. Be careful with **weighted measures** as they are **hard to interpret**
9. Agreement should be **above 0.8** to ensure data reliability (but depends on the case)

Bibliography
Where to Start Reading

TECHNISCHE UNIVERSITÄT DARMSTADT

Artstein, R./Poesio, M.: Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics* 34(4):555–596, 2008.

Artstein, R./Poesio, M.: Bias decreases in proportion to the number of annotators, In: *Proceedings of the 10th conference on Formal Grammar and the 9th Meeting on Mathematics of Language*, pp. 141–150, 2005.

Bennett, E.M./Alpert, R./Goldstein, A.C.: Communications through limited response questioning, *Public Opinion Quarterly* 18(3):303–308, 1954.

Carletta, J.: Assessing agreement on classification tasks: The kappa statistic, *Computational Linguistics* 22(2):249–254, 1996.

Cicchetti, D.V.: A new measure of agreement between rank ordered variables, In: *Proceedings of the American Psychological Association*, pp. 17–18, 1972.

Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* 70(4):213–220, 1968.

Cohen, J.: A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* 20(1):37–46, 1960.

Davies, M./Fleiss, J.L.: Measuring agreement for multinomial data, *Biometrics* 38(4):1047–1051, 1982.

Di Eugenio, B.: On the usage of Kappa to evaluate agreement on coding tasks, In: *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 441–444, 2000.

Di Eugenio, B./Glass, M.: The Kappa Statistic: A Second Look, *Computational Linguistics* 30(1):95–101, 2004.

Fleiss, J./Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement* 33(3):613–619, 1973.

Fleiss, J.L.: Measuring nominal scale agreement among many raters, *Psychological Bulletin* 76(5):378–381, 1971.

Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*, Thousand Oaks, CA: Sage Publications, 2004.

Landis, J.R./Koch, G.: The measurement of observer agreement for categorical data, *Biometrics* 33(1):159–174, 1977.

Neuendorf, K.A.: *The Content Analysis Guidebook*, Thousand Oaks, CA: Sage Publications, 2002.

Passonneau, R.J.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation, In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

Scott, W.A.: Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quaterly* 19(3):321–325, 1955.

Siegel, S./Castellan jr., N.J.: Nonparametric Statistics for the Behavioral Sciences, New York, NY: McGraw-Hill, 1988.

Join the Community!

DKPro Agreement

http://code.google.com/p/dkpro-statistics/

Announcements and discussion:

http://groups.google.com/group/dkpro-statistics-users

Download and issue tracker:

https://code.google.com/p/dkpro-statistics/

Project background:

https://www.ukp.tu-darmstadt.de/software/dkpro-statistics/