Measuring agreement between raters

by

Stian Lydersen

Presentation at RBUP, Oslo, 9 March 2016

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.

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. (Helvik et al. 2015)
- Psychiatric diagnosis based on Kiddie-SADS, based recorded telephone interview in the CAP (Hel-BUP) follow up study in Trondheim
- Rett’s-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:

<table>
<thead>
<tr>
<th>Level of measurement</th>
<th>Nominal / categorical</th>
<th>Ordinal</th>
<th>Interval and ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 raters</td>
<td>Cohen’s kappa</td>
<td>Cohen’s weighted kappa</td>
<td>Bland-Altman plots</td>
</tr>
<tr>
<td></td>
<td>ICC</td>
<td>ICC</td>
<td>ICC</td>
</tr>
<tr>
<td>&gt;2 raters</td>
<td>Fleiss’ kappa</td>
<td>Kendall’s coefficient of concordance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ICC</td>
<td>ICC</td>
<td>ICC</td>
</tr>
</tbody>
</table>
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 category \( i \) by rater 1 and category \( j \) by rater 2, respectively. An intuitive measure of agreement is the probability that the raters agree, which is

\[
P_{\text{agreement}} = p_{11} + p_{22} + \cdots + p_{cc}.
\]

But part of this agreement is due to chance. Suppose that rater 1 assigns to category \( i \) with probability \( p_i \) and rater 2 assigns to category \( j \) with probability \( p_j \), independently of rater 1. Then, Cohen’s probability of agreement by chance is given by

\[
P_{\text{chance}} = p_1 \cdot p_1 + p_2 \cdot p_2 + \cdots + p_c \cdot p_c.
\]

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

\[
kappa = \frac{P_{\text{agreement}} - P_{\text{chance}}}{1 - P_{\text{chance}}}.
\]

Example: Table 14.7:

Estimated agreement proportion:

\[
\hat{\rho} = \frac{21 + 17 + 15 + 1}{85} = \frac{54}{85} = 0.64
\]

Cohen’s probability of agreement by chance:

\[
\hat{\rho}_c = \frac{28 \times 33 + 38 \times 22 + 16 \times 29 + 3 \times 1}{85^2} = 0.31,
\]

Cohen’s kappa:

\[
k = \frac{0.64 - 0.31}{1 - 0.31} = 0.47.
\]

<table>
<thead>
<tr>
<th>Value of $\kappa$</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.20</td>
<td>Poor</td>
</tr>
<tr>
<td>0.21 – 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81 – 1.00</td>
<td>Very good</td>
</tr>
</tbody>
</table>

Interpretation of kappa values

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

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

$$SE(\kappa) = \frac{p_1(1-p_1)}{N(1-j)},$$

where $p_1$ is the marginal probability of category 1, $N$ is the total number of observations, and $j$ is the number of categories.

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

A 95% CI based on the data in Table 14.9 is (0.43, 0.82). Some software uses other formulae, see Lydersen (2012) and references therein.

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

Table 14.11: Symmetrical imbalance

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Healthy</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>50</td>
</tr>
</tbody>
</table>

$\kappa = 0.35$

Table 14.12: Asymmetrical imbalance

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Healthy</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

$\kappa = 0.44$

Kappa depends on the marginal distribution:

Inter-rater reliability assessment, the CAP (He-BUP) study, (Schei et al. 2015)

28 participants (drawn randomly) were scored by two raters.

<table>
<thead>
<tr>
<th></th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
</tr>
</tbody>
</table>

Cohen’s kappa = 0.50

<table>
<thead>
<tr>
<th></th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychotic</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
</tr>
</tbody>
</table>

Cohen’s kappa = 0.0

Table 14.7: Assessments of BS xerograms by two radiologists (from Boye et al., 1992).

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>Reign</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Suspected cancer</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Cancer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>24</td>
</tr>
</tbody>
</table>

Total 85
Cohens weighted kappa:
Weights the degree of agreement (distance from the diagonal)

Linear weighted kappa: \[ w_i = \frac{1 - \left| e_i - i \right|}{2 - i} \]
With 4 categories, the weights are 1 on the diagonal, and 2/3, 1/3 and 0 off the diagonal.

Quadratic weighted kappa: \[ w_i = \frac{1 - (e_i - i)^2}{(2 - i)^2} \]
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

<table>
<thead>
<tr>
<th></th>
<th>Unweighted</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 0 0 0</td>
<td>1 8/9 5/9 0</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>0 1 0 0</td>
<td>8/9 1 8/9 5/9</td>
</tr>
<tr>
<td>0 0 1 0</td>
<td>0 0 1 0</td>
<td>5/9 8/9 1 8/9</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>0 0 0 1</td>
<td>0 5/9 8/9 1</td>
</tr>
</tbody>
</table>

Linear versus quadratic weighted kappa?
- No clear advice in the literature
- For the case of equal marginal distributions, that is, \( n_{ii} = n_{jj} \) for all \( i \), then the quadratic weighted \( \tilde{\kappa} \) is equal to the intraclass correlation coefficient \( ICC \), described in Section 14.8, except for a term involving the factor \( 1/N \)

Table 14.8 Assessments of 85 xeromammograms by two radiologists

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Benign</td>
</tr>
<tr>
<td>Normal</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Benign</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Suspect cancer</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Cancer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>38</td>
</tr>
</tbody>
</table>

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

Alternative measures, two raters:
- Assuming independence between raters:
  - Cohen’s kappa (1960)
  - Scott’s pi (1955)
  - Benet’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
Measures which differ only in terms of calculating chance agreement:

**Cohen's kappa** (1960) uses the product of the marginals,
\[ \hat{\pi}_c = \frac{\hat{\pi}_{11} \hat{\pi}_{22}}{\hat{\pi}_{1+} \hat{\pi}_{2+}} \]

where \( \hat{\pi}_c = \frac{n_{11} + n_{22}}{n} \), and \( \hat{\pi}_{1+} = \frac{n_{11} + n_{12}}{n} \)

**Scott's pi** (1955) uses the squared average of the marginals,
\[ \hat{\pi}_c = \frac{1}{2} \left( \frac{\hat{\pi}_{1+} + \hat{\pi}_{2+}}{2} \right) \]

**Bennet's sigma** (1954) assumes a uniform marginal:
\[ \hat{\pi}_c = \frac{1}{c} \]

**Gwet’s gamma** (2001, 2008) (Also called Gwet’s AC₁):
\[ \hat{\pi}_c = \frac{1}{c-1} \sum \hat{\pi}_i \left( 1 - \hat{\pi}_i \right), \]

where \( \hat{\pi}_i = (\hat{\pi}_{ii} + \hat{\pi}_{i+})/2 \), \( \hat{\pi}_{ii} = \frac{n_{ii}}{n} \), and \( \hat{\pi}_{i+} = \frac{n_{i+}}{n} \)

When \( c = 2 \), the equation reduces to
\[ \hat{\pi}_c = 2 \hat{\pi}_1 \hat{\pi}_2 \]

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)

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

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

The inter-rater reliability measures (to be estimated) can be expressed as below. These expressions are definitional, since \( N_{12}^{HE} \) etc are not observed.

Gwet’s gamma:
\[ \hat{\pi}_c = \frac{\sum N_{ii}^{HE} - \sum N_{1i}^{HE} + \sum N_{i2}^{HE} + \sum N_{ij}^{HE}}{N} \]

Aickin’s alpha:
\[ \alpha = \frac{\sum N_{ii}^{HE}}{N} \]
Gwet’s $\gamma_1$: Green framed in numerator. All except crossed out in denominator.

Aickin’s $\alpha$: Green framed in numerator. All in denominator.

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:

$$\hat{\Delta} = \hat{p}_d + \hat{p}_h - 2\sqrt{\hat{p}_d \hat{p}_h}$$

Table 14.9

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>54</td>
<td>1</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>18</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
<td>25</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

$\hat{\kappa} = 0.635, \hat{\pi} = 0.627, \hat{\sigma} = 0.694, \hat{\gamma}_1 = 0.741, \Delta = 0.766$

Table 14.10

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>68</td>
<td>1</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>5</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

$\hat{\kappa} = 0.320, \hat{\pi} = 0.294, \hat{\sigma} = 0.694, \hat{\gamma}_1 = 0.805, \Delta = 0.766$

Table 14.11: Symmetrical imbalance

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Disease</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>50</td>
<td>10</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>30</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

$\kappa = 0.348, \hat{\pi} = 0.341, \hat{\sigma} = 0.400, \hat{\gamma}_1 = 0.450, \Delta = 0.417$

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

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Disease</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>30</td>
<td>0</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>40</td>
<td>0</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>40</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

$\kappa = 0.444, \hat{\pi} = 0.394, \hat{\sigma} = 0.400, \hat{\gamma}_1 = 0.446, \Delta = 0.700$ (or 0.585)

Gwet’s gamma is paradox-resistant (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.

Comparisons of measures for 2 raters:

(Anto, Lopez, & Benavente 2011) compare measures in terms of their ability to estimate the systematic agreement proportion. Hence, the construct (estimand) is $\Delta (?)$. Recommend Bennett’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”!

Categorical data: Generalizations to more than two raters

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

More than two raters: Dichotomous data

Example:

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.

Software:
This free software turned out to have some errors:

We used this commercial software:
www.agreestat.com
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.

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

\[ p_y = P(\text{Case no } i \text{ is classified as } "1" \text{ by rater } j) \]

\[ \log \left( \frac{p_y}{1-p_y} \right) = \beta_0 + b_i + c_j \]

where

- \( b_i \sim N(0, \sigma_{b_i}^2) \) is the random effect of case (subject) number \( i \)
- \( c_j \sim N(0, \sigma_{c_j}^2) \) is the random effect of rater \( j \)
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
```

“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.”

More than two raters: Ordinal data

Ordinal measurement, more than two raters:


Rets 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 \( \rho \) 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)

Wave 1, 19 nurses:

\( W = 0.822 \)

Wave 2, 12 nurses:

\( W = 0.844 \)

Two raters, dichotomous data: Positive and negative agreement
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

<table>
<thead>
<tr>
<th>Test result</th>
<th>Disease status</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td></td>
<td>Non-disease</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{a}{a+b} \), Specificity = \( \frac{d}{c+d} \)

Continuous data: Intraclass correlation coefficient (ICC)

Example revisited: The CAP (Hei-BUP) study

<table>
<thead>
<tr>
<th></th>
<th>Anxiety</th>
<th>Psychotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>Cohen’s kappa=0.50</td>
<td>Cohen’s kappa=0.0</td>
</tr>
<tr>
<td>Rater 2</td>
<td>Positive agreement: 0.62</td>
<td>Positive agreement: 0.0</td>
</tr>
<tr>
<td></td>
<td>Negative agreement: 0.88</td>
<td>Negative agreement: 0.98</td>
</tr>
</tbody>
</table>

Schei et al. (2015)
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.

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.

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.

Case 1:
One-way random effect model

where

\( X_{ij} \) is rating number \( j \) on subject number \( i \),
\( b_i \sim N(0, \sigma_{b_i}^2) \) is the random effect of subject number \( i \),
\( w_{ij} \sim N(0, \sigma_{w_{ij}}^2) \) is a residual term.

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

The correlation between two ratings \( X_{ij} \) and \( X_{ij_0} \) on subject number \( i \) is

\[
ICC_i = \frac{\sigma_{e}^2}{\sigma_{j}^2 + \sigma_{e}^2}
\]
In Case 1, \( w_j \) includes a rater effect and an error term.

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

\[
X_{ij} = \mu + c_j + e_{ij},
\]

where

- \( c_j \) is the effect of rater \( j \)
- \( e_{ij} \) is the residual random error.

Case 2: \( e_{ij} \sim N(0, \sigma^2_e) \)

Case 3, \( e_{ij} \) is a fixed effect with constraint \( \sum_{j=1}^{n} e_{ij} = 0 \).

The correlation between two ratings \( X_{ij} \) and \( X_{ij'} \) on subject number \( i \) is

\[
ICC = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_C + \sigma^2_E}.
\]

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 measurements as a function of their mean, commonly termed a Bland-Altman plot.

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.

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.
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  

Linear model with crossed random effects of individual and rater

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

$$X_{ij} = \beta_0 + \beta_{time_i} + \beta_{time_j} + b_i + c_j + \epsilon_{ij}$$

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

(Show results from Word document)

There are 3 variance components:

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

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

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

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

The average Pearson correlation between the raters was 0.63.

Effect of rating 2 versus rating 1 on same individual?

$$X_{ij} = \beta_0 + \beta_{time_i} + \beta_{time_j} + \beta_{rating_2} + b_i + c_j + \epsilon_{ij}$$

(Show results from word document)

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.

<table>
<thead>
<tr>
<th>Second rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>First rater</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>196</td>
</tr>
</tbody>
</table>

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
- Rater: 9.789 = 3.129
- Residual: 27.120 = 5.208

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

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

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

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.


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


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


Henning et al (2016) revisited. 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 ± 0.077², including the variance between the rated patients (0.627 ± 0.792²), plus the variance due to the raters (0.00212 ± 0.046²) and the residual variance (0.179 ± 0.371²). The interrater reliability (ICC) estimate was 0.816.