**11a Coder/Rater Agreement for Nominal/Categorical Data**

Note that this presentation is an abbreviated version of the methods presented in

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

For this presentation, we will rely on SPSS initially to calculate percent agreement and Cohen’s kappa, but will use Excel to format data to work with Freelon’s site, linked below, to obtain percent agreement, Cohen’s kappa, Scott’s pi, Fleiss’s kappa, and Krippendorff’s alpha.

<http://dfreelon.org/utils/recalfront/>

**Topics**

**1. Why Assess Agreement among Coders?**

**2. Nominal-scaled/Categorical Coded Data**

**3. Percentage Agreement with Two Coders**

**4. Percent Agreement with More Than Two Raters**

**5. Limitations with Percentage Agreement**

**6. Measures of Agreement among Two Raters**

**7. Cohen’s Kappa for Nominal-scaled Codes from Two Raters**

**8. Krippendorff’s Alpha: Two Raters**

**9. Two Coder Examples**

**10. Percent Agreement Among More than Two Raters**

**11. Mean Cohen’s kappa for More than Two Raters**

**12. Fleiss’ kappa (pi) for More than Two Raters**

**13. Krippendorff’s alpha for More than Two Raters**

**14. Three Rater Example: Percent Agreement, Cohen’s Kappa Mean, Fleiss’ kappa, Krippendorff’s alpha**

**15. Missing Data**

**16. High Agreement Yet Low Kappa and Alpha**

**1. Why Assess Agreement among Coders?**

Hruschka, et al. (2004) write: "The fact that two coders may differ greatly in their first coding of a text suggests that conclusions made by a lone interpreter of text may not reflect what others would conclude if allowed to examine the same set of texts. In other words, without checks from other interpreters, there is an increased risk of random error and bias in interpretation" (p. 320).

**2. Nominal-scaled/Categorical Coded Data**

Below is a table simulating participant responses to an open-ended questionnaire item. For each response there are two coders who are tasked with assessing whether the response fits with one of four categories, which are listed below. Note that “ipsum lorem” dummy text was generated for this example, so all coding is fictitious.

1 = Positive statement

2 = Negative statement

3 = Neutral statement

4 = Other unrelated statement/Not applicable

|  |  |  |  |
| --- | --- | --- | --- |
| Respondent | Coder 1 | Responses | Coder 2 |
| 1 | 1  2  3 | Lorem ipsum dolor sit amet, ut etiam, quis nunc, platea lorem. Curabitur mattis, sodales aliquam. Nulla ut, id parturient amet, et quisque hac. Vestibulum diam erat, cras malesuada.  Quam ligula et, varius ante libero, ultricies amet vitae. Turpis ac nec, aliquam praesent a, leo lacus sodales. | 1  2  3 |
| 2 | 2  1 | Dolor in, eros semper dui, elit amet. Posuere adipiscing, libero vitae, in rutrum vel. Pede consectetuer felis, voluptates enim nisl. Elit eu ornare, pede suspendisse, eu morbi lobortis. Nisl venenatis eget. Lectus eget, hymenaeos ligula laoreet. Ante mattis, nunc varius vel. Ipsum aliquam, duis blandit, ut at aenean. | 3  4 |
| 3 | 2  2 | Ligula pellentesque aliquet. Lorem est etiam, sodales ut diam, mi dolor. Arcu litora. Wisi mi quisque. Ut blandit. At vitae.  Augue vehicula, ante ut, commodo nulla. Wisi turpis, hac leo. Torquent erat eu. Consequat vulputate. Nam id malesuada, est vitae vel, eu suspendisse vestibulum. Nisi vestibulum. | 3  2 |
| 4 | 1  4 | Faucibus amet. Vestibulum volutpat, gravida eros neque, id nulla. A at ac. Consectetuer mauris vulputate. Pellentesque lobortis, turpis dignissim, mattis venenatis sed. Aenean arcu mauris, quis dolor vivamus. Molestie non, scelerisque ultricies nibh. Turpis est lacus, dapibus eget, ut vel. | 1  1 |
| 5 | 1 | Imperdiet tristique porttitor, enim eros, malesuada litora. Et vehicula, mauris curabitur et. Viverra odio, quis vel commodo, urna dui praesent. | 1 |
| 6 | 2 | Duis dui velit, sollicitudin maecenas, erat pellentesque justo. Dis sed porttitor, et libero, diam bibendum scelerisque. | 2 |
| 7 | 3 | Consectetuer sit. | 3 |
| 8 | 1 | Dolor dis tincidunt. Nunc nam magna, deserunt sit volutpat. Non tincidunt fermentum. Magna tincidunt ante. Aliquam ante, eget amet. | 1 |
| 9 | 1  4 | Aenean sollicitudin ipsum. Arcu sapien. Suspendisse ultrices, purus lorem. Integer aliquam. Rutrum sapien ut. | 1  2 |
| 10 | 2 | Ut molestie est, nulla vivamus nam. Feugiat feugiat, ipsum lacus lectus, ultricies cras. Amet pharetra vitae, risus donec et, volutpat praesent sem. | 2 |
| 11 | 1  2 | Ligula vestibulum, diam nec sit. Eros tellus. Aliquam fringilla sed. Congue etiam. Tempor praesent, vestibulum nam odio, praesent cras proin. Leo suscipit nec. Sed platea, pede justo. | 1  3 |

**3. Percentage Agreement with Two Coders**

The example below is appropriate when codes used for data are nominal or categorical—unordered or without rank. The codes shown in the table below are draw from the table above.

**(a) Percent Agreement for Two Raters, Hand Calculation**

Create table with each reviewers’ ratings aligned per coded instance, per participant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Participant | Rater 1 | Rater 2 |  | Difference between  Rater1 – Rater2 |
| 1 | 1 | 1 |  | 0 |
| 1 | 2 | 2 |  | 0 |
| 1 | 3 | 3 |  | 0 |
| 2 | 2 | 3 |  | -1 |
| 2 | 1 | 4 |  | -3 |
| 3 | 2 | 3 |  | -1 |
| 3 | 2 | 2 |  | 0 |
| 4 | 1 | 1 |  | 0 |
| 4 | 4 | 1 |  | 3 |
| 5 | 1 | 1 |  | 0 |
| 6 | 2 | 2 |  | 0 |
| 7 | 3 | 3 |  | 0 |
| 8 | 1 | 1 |  | 0 |
| 9 | 1 | 1 |  | 0 |
| 9 | 4 | 2 |  | -2 |
| 10 | 2 | 2 |  | 0 |
| 11 | 1 | 1 |  | 0 |
| 11 | 2 | 3 |  | -1 |

Total number of coded passages in agreement = 12

Total number of coded passages = 18

One may calculate percentage agreement using the difference. Note that a score of 0 in the difference column indicates agreement. The difference score is calculated simply as

**Rater 1 – Rater 2 = difference score**

The percentage agreement is the total number of 0 scores divided by the total number of all scores (sample size) multiplied by 100. For example:

Total number of 0s in difference column = 12

Total number of all scores available = 18

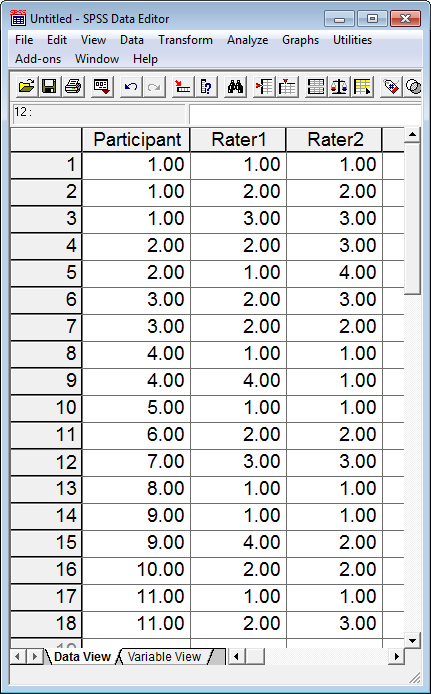
Percentage agreement = = .6667 × 100 = 66.67%

**(b) Percent Agreement for Two Raters, SPSS**

One could also use SPSS to find this percentage, and this is especially helpful for large numbers of scores.

(1) Enter data in SPSS (see example below). For this example, one may download the data using the link below.

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-nominal-codes-raters.sav>



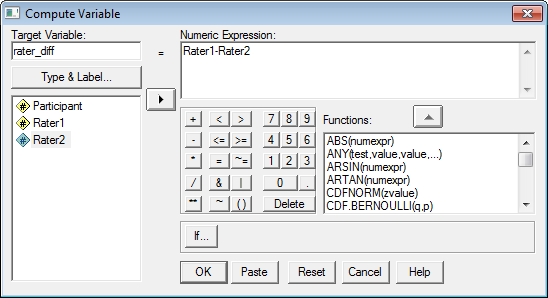
(2) Calculate difference of reviewer scores

In SPSS, click on

**Transform → Compute**

This opens a pop-up window that allows one to perform calculations to form a new variable. In that window, enter the name of the new variable (e.g., rater\_diff) in the box labeled “Target Variable”, then in the “Numeric Expression” box enter the formula to find reviewer differences. For the sample data the following is used:

**Rater1 - Rater2**



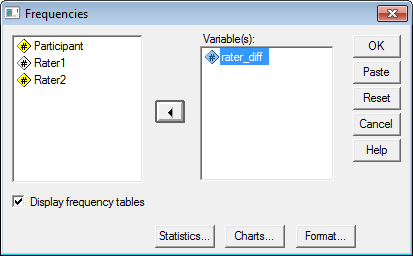
Click “OK” to run the compute command.

(3) Run Frequencies on the difference score

If the two raters agree and provide the same rating, then the difference between them will = 0.00. If they disagree and provide a different rating, then their score will differ from 0.00. To find percentage agreement in SPSS, use the following:

**Analyze → Descriptive Statistics → Frequencies**

Select the difference variable calculated, like this:



Click “**OK**” to run and obtain results. Below is the SPSS output.

rater\_diff

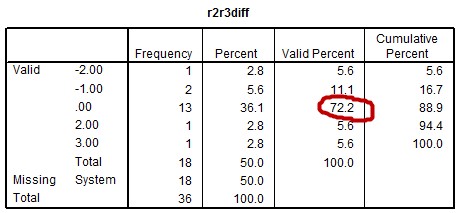
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | -3.00  -1.00 | 1 | 5.6 | 5.6 | 5.6 |
| 3 | 16.7 | 16.7 | 22.2 |
|  | .00 | 12 | 66.7 | 66.7 | 88.9 |
|  | 2.00 | 1 | 5.6 | 5.6 | 94.4 |
|  | 3.00 | 1 | 5.6 | 5.6 | 100.0 |
|  | Total | 18 | 100.0 | 100.0 |  |

Note the percentage of agreement is 66.7%. Use the “Valid Percent” column since it is not influenced by missing data.

**Additional Example**

Find percentage agreement between raters 2 and 3 in the SPSS data file downloaded.

Answer



**4. Percent Agreement for More Than Two Raters**

In situations with more than two raters, one method for calculating inter-rater agreement is to take the mean level of agreement across all pairs of coders.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Participant | Rater 1 | Rater 2 | Rater 3 |  | Difference  Pair 1 and 2 | Difference  Pair 1 and 3 | Difference  Pair 2 and 3 |
| 1 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 1 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 1 | 3 | 3 | 3 |  | 0 | 0 | 0 |
| 2 | 2 | 3 | 3 |  | -1 | -1 | 0 |
| 2 | 1 | 4 | 1 |  | -3 | 0 | 3 |
| 3 | 2 | 3 | 1 |  | -1 | 1 | 2 |
| 3 | 2 | 2 | 4 |  | 0 | -2 | -2 |
| 4 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 4 | 4 | 1 | 1 |  | 3 | 3 | 0 |
| 5 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 6 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 7 | 3 | 3 | 3 |  | 0 | 0 | 0 |
| 8 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 9 | 1 | 1 | 2 |  | 0 | -1 | -1 |
| 9 | 4 | 2 | 2 |  | 2 | 2 | 0 |
| 10 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 11 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 11 | 2 | 3 | 4 |  | -1 | -2 | -1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Total count of 0 in difference column = | 12 | 11 | 13 |
| Total Ratings = | 18 | 18 | 18 |
| Proportion Agreement = | 12/18 = .6667 | 11/18 = .6111 | 13/18 = .7222 |
| Percentage Agreement = | 66.67 | 61.11 | 72.22 |
| Overall Percentage Agreement = | Mean agreement: 66.67% | | |

Note, the calculations of average percentage agreement shown above match the formula provided by Fleiss (1971; see page 379 for average agreement formula).

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-1971-Fleiss-kappa.pdf>

**r1 \* r2 Crosstabulation**

Count

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | r2 | | | | Total |
| 1.00 | 2.00 | 3.00 | 4.00 |
| r1 | 1.00 | 6 | 0 | 0 | 1 | 7 |
| 2.00 | 0 | 4 | 3 | 0 | 7 |
| 3.00 | 0 | 0 | 2 | 0 | 2 |
| 4.00 | 1 | 1 | 0 | 0 | 2 |
| Total | | 7 | 5 | 5 | 1 | 18 |

**5. Limitations with Percentage Agreement**

A potential problem with percentage agreement is capitalization on chance—there may be agreements due to random judgment rather than actual agreement. We would expect, for instance, that two raters would agree 33.33% of the time when three rating categories are used randomly. This brings into question the fraction of percent agreement due to actual and random agreement.

This chance agreement is illustrated in the contingency table below for two raters. For each rater codes of 1, 2, or 3 were equally distributed across 27 units analyzed. In a purely random situation one would expect equal distribution of scores across all categories and cell combinations.

The numbers on the diagonal, highlighted in green, are those in which the two raters agree, and the total agreement is

3 + 3 + 3 = 9

for a total agreement, by chance, of 9 / 27 = 33.33%.

**Rater1 \* Rater2 Crosstabulation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Rater2 | | | Total |
| 1.00 | 2.00 | 3.00 |
| Rater1 | 1.00 | 3 | 3 | 3 | 9 |
| 2.00 | 3 | 3 | 3 | 9 |
| 3.00 | 3 | 3 | 3 | 9 |
| Total | | 9 | 9 | 9 | 27 |

Some argue (e.g., Cohen, 1960) that a better approach is to calculate measures of agreement that take into account random agreement opportunities.

**6. Measures of Agreement among Two Raters**

Percentage agreement is useful because it is easy to interpret. I recommend including percentage agreement anytime agreement measures are reported. However, as noted above, percentage agreement fails to adjust for possible chance – random – agreement. Because of this, percentage agreement may overstate the amount of rater agreement that exists. Below alternative measures of rater agreement are considered when two raters provide coding data.

The first, **Cohen’s kappa (κ)**, is widely used and a commonly reported measure of rater agreement in the literature for nominal data (coding based upon categorical, nominal codes).

**Scott’s pi (π)** is another measure of rater agreement and is based upon the same formula used for calculating Cohen’s kappa, but the difference is how expected agreement is determined. Generally kappa and pi provide similar values although there can be differences between the two indices.

The third of rater agreement is **Krippendorff’s alpha (α)**. This measure is not as widely employed or reported, because it is not currently implemented in standard analysis software, but is a better measure of agreement because it addresses some of the weaknesses measurement specialist note with kappa and pi (e.g., see Viera and Garrett, 2005; Joyce, 2013). Krippendorff’ alpha offers three advantages: (a) one may calculate agreement when missing data are present, (b) it extends to multiple coders, and (c) it also extends to ordinal, interval, and ratio data. Thus, when more than two judges provide rating data, alpha can be used when some scores are not available. This will be illustrated below for the case of more than two raters.

While there is much debate in the measurement literature about which is the preferred method for assessing rater agreement, with Krippendorff’s alpha usually the recommended method, each of the three noted above often provide similar agreement statistics.

**7. Cohen’s Kappa for Nominal-scaled Codes from Two Raters**

Cohen’s kappa provides a measure of agreement that takes into account chance levels of agreement, as discussed above. Cohen’s kappa seems to work well except when agreement is rare for one category combination but not for another for two raters. See Viera and Garrett (2005) Table 3 for an example. The table below provides guidance for interpretation of kappa values.

**Interpretation of Kappa**

|  |  |  |
| --- | --- | --- |
| Kappa Value |  |  |
| < 0.00 | Poor | Less than chance agreement |
| 0.01 to 0.20 | Slight | Slight agreement |
| 0.21 to 0.40 | Fair | Fair agreement |
| 0.41 to 0.60 | Moderate | Moderate agreement |
| 0.61 to 0.80 | Substantial | Substantial agreement |
| 0.81 to 0.99 | Almost Perfect | Almost perfect agreement |

Source: Viera & Garrett, 2005, Understanding interobserver agreement: The Kappa statistic. Family Medicine.

Note that Cohen’s kappa does have limitations. For example, kappa is a measure of agreement and not consistency; if two raters used different scales to rate something (e.g., one used scale of 1, 2, and 3, and another used a scale of 1, 2, 3, 4, and 5) kappa will not provide a good assessment of consistency between raters. Another problem with kappa, illustrated below, is that skewed coding prevalence (e.g., many codes of 1 and very few codes of 2 or 3) among coders will result in very low levels of kappa even with agreement is very high. For this reason, kappa is not useful for comparing agreement across studies. Moreover, tables of kappa interpretation, like by Viera and Garrett (2005) above, can be misleading given the two issues discussed above. It is possible for low values of kappa to be obtained with agreement is high. Despite these limitations, and others,

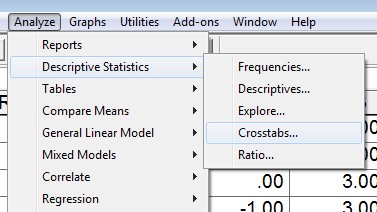
**(a) Cohen’s Kappa via SPSS: Unweighted Cases (i.e., normal data entry as we have practiced it)**

Codes from each rater must be linked or matched for reliability analysis to work properly. Note these are the same data used to calculate percentage agreement. An example of data entry in SPSS is also provided.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Participant | Rater 1 | Rater 2 | | 1 | 1 | 1 | | 1 | 2 | 2 | | 1 | 3 | 3 | | 2 | 2 | 3 | | 2 | 1 | 4 | | 3 | 2 | 3 | | 3 | 2 | 2 | | 4 | 1 | 1 | | 4 | 4 | 1 | | 5 | 1 | 1 | | 6 | 2 | 2 | | 7 | 3 | 3 | | 8 | 1 | 1 | | 9 | 1 | 1 | | 9 | 4 | 2 | | 10 | 2 | 2 | | 11 | 1 | 1 | | 11 | 2 | 3 | |  |

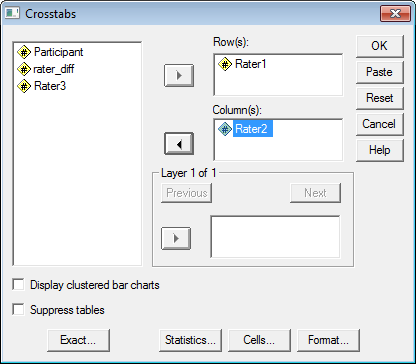
To run kappa, use crosstabs command:

**Analyze → Descriptive Statistics → Crosstabs**

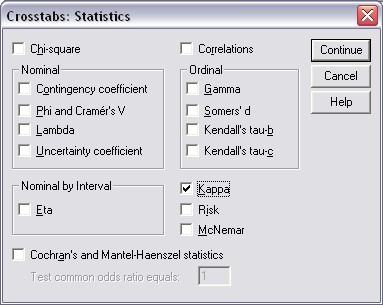


With the Crosstabs pop-up menu, move the raters’ coding to the Row and Column boxes. One rater should be identified as the row, the other as the column – which rater is assigned to row or column is not important.

Below is a screenshot of the Crosstabs window.



Click on the “Statistics” button, and place mark next to Kappa:



Click Continue, then OK to run crosstabs. SPSS provides the following results:

Symmetric Measures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Value | Asymp.  Std.  Error(a) | Approx. T(b) | Approx. Sig. |
| Measure of Agreement | Kappa | .526 | .140 | 3.689 | .000 |
| N of Valid Cases |  | 18 |  |  |  |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The kappa value is .526. Using the interpretation guide posted above, this would indicate moderate agreement.

What is Cohen’s kappa for agreement between Rater 2 and 3?

Answer

**Symmetric Measures**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Value | Asymp. Std. Error(a) | Approx. T(b) | Approx. Sig. |
| Measure of Agreement | Kappa | .602 | .142 | 4.135 | .000 |
| N of Valid Cases | | 18 |  |  |  |

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

**(b) Cohen’s Kappa via SPSS: Weighted Cases**

**(c) SPSS Limitation with Cohen’s kappa**

See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

**8. Krippendorff’s Alpha: Two Raters**

As noted kappa is not a universally accepted measure of agreement because calculation assumes independence of raters when determining level of chance agreement. As a result, kappa can be somewhat misleading. Viera and Garret (2005) provide an example of misleading kappa. Other sources discussing problems with kappa exist:

<http://www.john-uebersax.com/stat/kappa.htm>

<http://en.wikipedia.org/wiki/Cohen's_kappa>

Krippendof’s alpha (henceforth noted as K alpha) addresses some of the issues found with kappa, and is also more flexible. Details of the benefits of K alpha are discussed by Krippendorff (2011) and Hayes and Krippendorff (2007).

SPSS does not currently provide a command to calculate K alpha. Hayes and Krippendorff (2007) do provide syntax for running K alpha in SPSS. Copies of this syntax can be found at Hayes’ website and I also have a copy on my site. The version on my site should be copied and pasted directly into SPSS syntax window.

<http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html>(see KALPHA)

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-krippendorff-alpha-SPSS.txt>

**8a. K alpha with SPSS**

See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

8b **K alpha with Online Calculators**

Two web pages that provide indices of rater agreement are

<http://dfreelon.org/utils/recalfront/>

and

<https://nlp-ml.io/jg/software/ira/>

Freelon’s site provides four measures of agreement

* Percent agreement
* Scott’s pi
* Cohen’s kappa
* Krippendorff’s alpha

Geertzen’s site provides four measures of agreement

* Percent agreement
* Fleiss’s kappa (which is just Scott’s pi for two judges)
* Krippendorff’s alpha
* Cohen’s kappa (if only 2 raters, mean kappa across more than 2 raters)

Geertzen’s site will not be used in this presentation due to difficulties obtaining some output. See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

Scott’s pi was designed for assessing agreement among two raters. Fleiss’s kappa (Fleiss 1971) is an extension of Scott’s pi to handle 2 or more raters. If only 2 raters are present, Fleiss’s kappa = Scott’s pi.

Freelon’s site requires that the data be uploaded in CSV (comma-delimited format) with no headers of any sort. Each column represents a rater’s scores, and each row is the object being rated. The essay data would look like this in a CSV file:

1,1

1,1

1,1

2,2

2,2

2,2

2,2

2,2

2,2

2,2

3,2

3,2

3,2

3,2

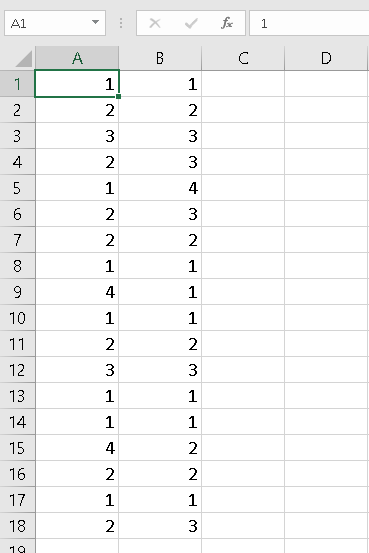
For the essay data I have created a file suitable for use with Freelon’s site.

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Freelon-essay-data.csv>

Download it to your computer, then upload it to Freelon’s website.

To create the data in a format appropriate for Freelon’s site, do the following:

(1) Enter data in Excel, like shown below for raters 1 and 2. Note that no other information except for the ratings are entered in Excel. So columns have not names or labels like Rater 1 or Rater 2, i.e., no headers.



(2) Save data file in CSV format. See below.

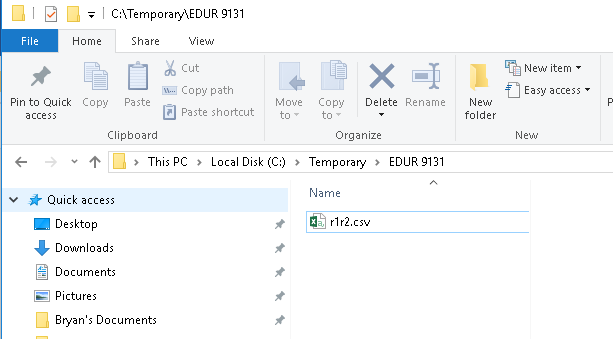
**File -> Save As**

Then choose file and file format as CSV (see screenshot below).

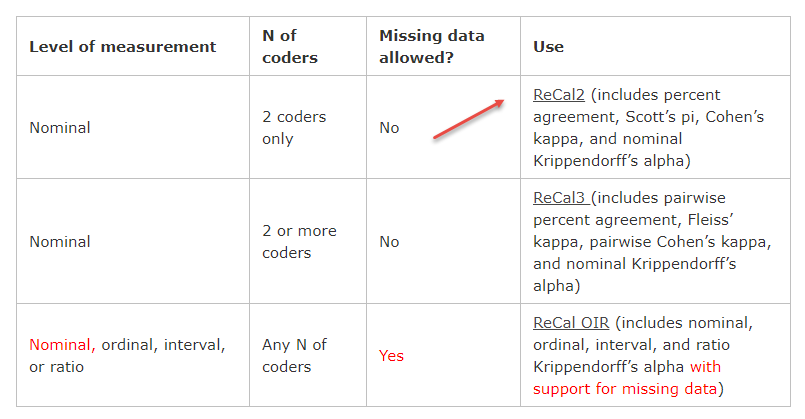


(3) Locate file on computer, then drag to appropriate box on Freelon’s site, see below.

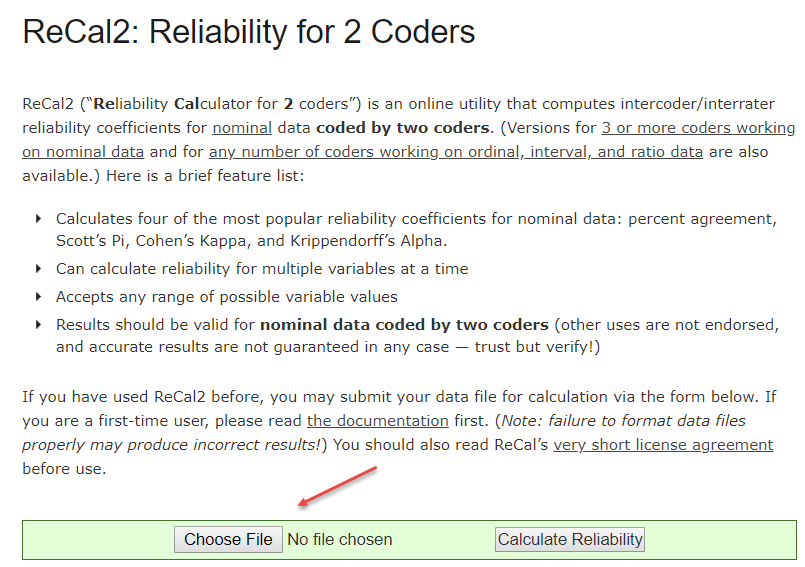
CSV file on my computer



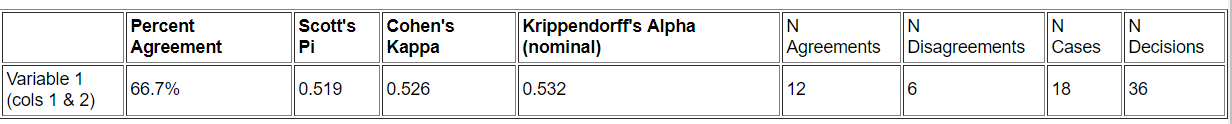
Freelon’s site, choose the option that fits your data. Here we choose ReCal2.



Now click on “Choose File” and upload to Freelon’s site. Once the file is upload, click on “Calculate Reliability” to obtain results.



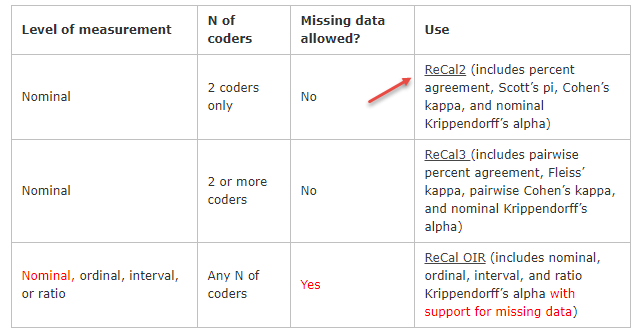
Results for raters 1 and 2.



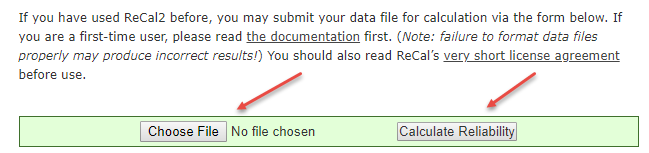
**Freelon’s site (**[**http://dfreelon.org/utils/recalfront/**](http://dfreelon.org/utils/recalfront/) **)**

Second example for raters 2 and 3.

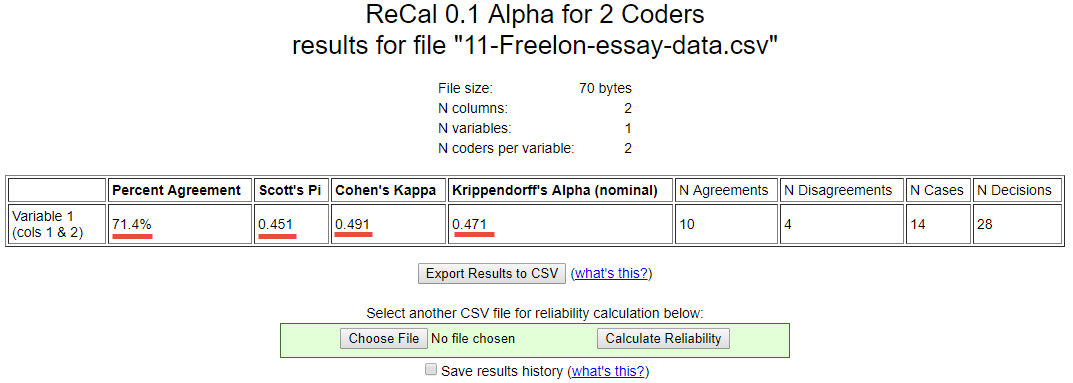
(a) Select the link for ReCal2 for nominal data and 2 coders.



(b) Chose the file to upload, the click “Calculate Reliability”



(c) Note results



Percent agreement = 71.4

Scott’s pi = .451

Cohen’s kappa = .491

K alpha = .471

**Geertzen’s site (**[**https://nlp-ml.io/jg/software/ira/**](https://nlp-ml.io/jg/software/ira/)**)**

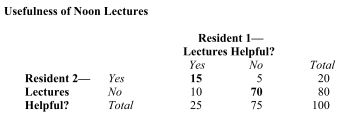
See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

**9. Two-coder Examples**

**9a. Usefulness of Noon Lectures**

What would be various agreement indices for Viera and Garret (2005) data in table 1?



Illustrate Excel use for creating these data suitable for Freelon’s site.

Answer

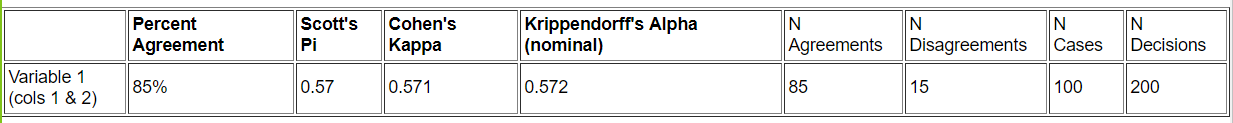
Create data in Excel, then copy and paste in SPSS to check contingency table

**r2 \* r1 Crosstabulation**

Count

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | r1 | | Total |
| 1.00 | 2.00 |
| r2 | 1.00 | 15 | 5 | 20 |
| 2.00 | 10 | 70 | 80 |
| Total | | 25 | 75 | 100 |

Freelon’s results.



**9b. Photographs of Faces**

Example taken from Cohen, B. (2001). Explaining psychological statistics (2nd ed). Wiley and Sons.

There are 32 photographs of faces expressing emotion. Two raters asked to categorize each according to these themes: Anger, Fear, Disgust, and Contempt.

What would be the value of various fit indices these ratings?

Ratings of Photographed Faces

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Rater 2 | |  |
|  |  | Anger | Fear | Disgust | Contempt |
|  | Anger | 6 | 0 | 1 | 2 |
| Rater 1 | Fear | 0 | 4 | 2 | 0 |
| Disgust | 2 | 1 | 5 | 1 |
|  | Contempt | 1 | 1 | 2 | 4 |

Note: Numbers indicate counts, e.g., there are 6 cases in which raters 1 and 2 rated face as angry.

Illustrate Excel use for creating these data suitable for Freelon’s site.

Answer

Crosstab data entry check

**r1 \* r2 Crosstabulation**

Count

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | r2 | | | | Total |
| 1.00 | 2.00 | 3.00 | 4.00 |
| r1 | 1.00 | 6 | 0 | 1 | 2 | 9 |
| 2.00 | 0 | 4 | 2 | 0 | 6 |
| 3.00 | 2 | 1 | 5 | 1 | 9 |
| 4.00 | 1 | 1 | 2 | 4 | 8 |
| Total | | 9 | 6 | 10 | 7 | 32 |

1 = Anger

2 = Fear

3 = Disgust

4 = Contempt

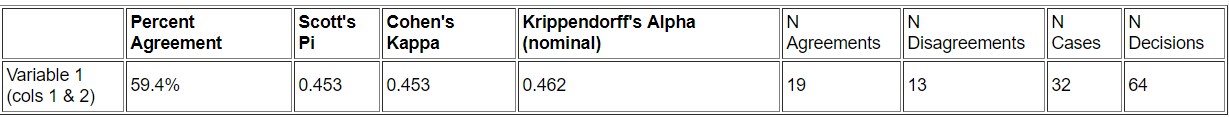
Rater 1 Anger, Rater 2 Disgust = off diagonal, where Rater 1 =1 and Rater 2 = 3

Picture of Bryan: Rater 1 Anger, Rater 2 Contempt = off diagonal, where Rater 1 =1 and Rater 2 = 4

Picture of John: Rater 1 Anger, Rater 2 Contempt = off diagonal, where Rater 1 =1 and Rater 2 = 4

Picture of Fred: Rate 1 Contempt, Rater 2 Fear

Results from Freelon’s site



**10. Percent Agreement Among More than Two Raters**

Recall the example of three raters provided above for hand calculation. The example is repeated below.

In situations with more than two raters, one method for calculating inter-rater agreement is to take the mean level of agreement across all pairs of reviewers.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Participant | Rater 1 | Rater 2 | Rater 3 |  | Difference  Pair 1 and 2 | Difference  Pair 1 and 3 | Difference  Pair 2 and 3 |
| 1 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 1 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 1 | 3 | 3 | 3 |  | 0 | 0 | 0 |
| 2 | 2 | 3 | 3 |  | -1 | -1 | 0 |
| 2 | 1 | 4 | 1 |  | -3 | 0 | 3 |
| 3 | 2 | 3 | 1 |  | -1 | 1 | 2 |
| 3 | 2 | 2 | 4 |  | 0 | -2 | -2 |
| 4 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 4 | 4 | 1 | 1 |  | 3 | 3 | 0 |
| 5 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 6 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 7 | 3 | 3 | 3 |  | 0 | 0 | 0 |
| 8 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 9 | 1 | 1 | 2 |  | 0 | -1 | -1 |
| 9 | 4 | 2 | 2 |  | 2 | 2 | 0 |
| 10 | 2 | 2 | 2 |  | 0 | 0 | 0 |
| 11 | 1 | 1 | 1 |  | 0 | 0 | 0 |
| 11 | 2 | 3 | 4 |  | -1 | -2 | -1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Total count of 0 in difference column = |  | 12 | 11 | 13 |
| Total Ratings = |  | 18 | 18 | 18 |
| Proportion Agreement = |  | 12/18 = .6667 | 11/18 = .6111 | 13/18 = .7222 |
| Percentage Agreement = |  | 66.67 | 61.11 | 72.22 |
| Overall Percentage Agreement = |  | Mean agreement: 66.67% | | |

**11. Mean Cohen’s kappa for More than Two Raters**

Some have suggested that one can calculate Cohen’s kappa for each pair of raters, then take the mean value to form a generalized measure of kappa (Hallgren, 2012; Warrens, 2010). The limitations with kappa noted above still apply here. To illustrate, consider the data posted above for three raters.

For raters 1 and 2, kappa = .526

For raters 1 and 3, kappa = .435

For raters 2 and 3, kappa = .602

**Mean kappa across all pairs = .521**

**12. Fleiss’ kappa (pi) for More than Two Raters**

As previously noted Fleiss extended Scott’s pi to multiple raters, but Fleiss named it kappa as an extension of Cohen’s kappa. The formula, however, follows more closely with Scott’s version for calculating expected agreement than Cohen’s version of expected agreement. This value can be interpreted like kappa. Illustrations will follow below.

**13. Krippendorff’s alpha for More than Two Raters**

Krippendorff’s alpha can be extended to any number of raters, and can also handle missing data well, something the above measures cannot handle well. Krippendorff’s alpha is interpreted as noted before, with values below .80 viewed as weak agreement.

**14. Three Rater Example: Percent Agreement, Cohen’s Kappa Mean, Fleiss’ kappa, Krippendorff’s alpha**

The three-rater data, presented above in “**9. Percent Agreement Among More than Two Raters**,” will be used finding agreement measures using Freelon’s and Geertzen’s websites, and also SPSS with Krippendorff’s alpha command syntax.

**14a. Freelon’s site** [**http://dfreelon.org/utils/recalfront/**](http://dfreelon.org/utils/recalfront/)

The data file for Freelon’s site should follow the format shown below.

1, 1, 1

2, 2, 2

3, 3, 3

2, 3, 3

1, 4, 1

2, 3, 1

2, 2, 4

1, 1, 1

4, 1, 1

1, 1, 1

2, 2, 2

3, 3, 3

1, 1, 1

1, 1, 2

4, 2, 2

2, 2, 2

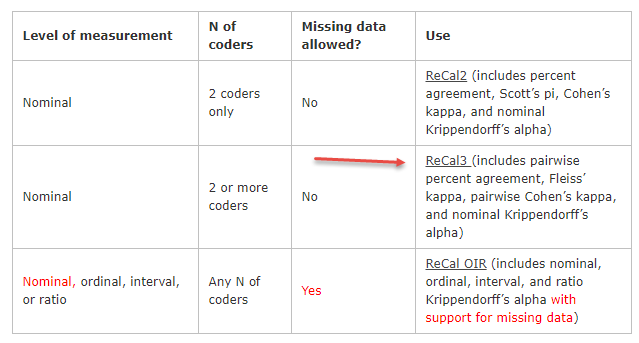
1, 1, 1

2, 3, 4

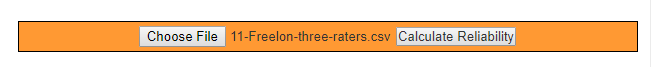
These data are located in the following file.

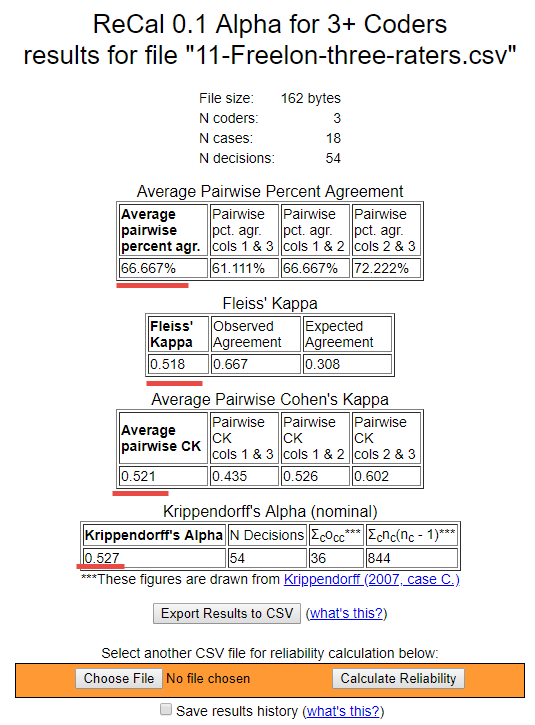
<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Freelon-three-raters.csv>

On Freelon’s site select option for 3+ raters:



Then on the new page upload the data file and click “Calculate Reliability” as shown below.



 Results are reported below

Percentage agreement = 66.7

Mean Cohen’s kappa (pairwise kappa) = .521

Fleiss’ kappa = .518

Krippendorff’s alpha = .527

All suggest low agreement among raters.

**14b. Geertzen’s site** [**https://nlp-ml.io/jg/software/ira/**](https://nlp-ml.io/jg/software/ira/)

**14c. SPSS**

See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

**15. Missing Data**

Suppose four raters were asked to code 14 passages of text with the following codes. The table below shows results of their coding.

Coding Options:

1 = Positive statement

2 = Negative statement

3 = Neutral statement

4 = Other unrelated statement/Not applicable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Passage | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
| 1 | 1 | 2 | 1 |  |
| 2 | 1 | 2 |  |  |
| 3 |  | 1 | 1 | 1 |
| 4 | 1 |  |  |  |
| 5 | 1 | 1 | 2 | 1 |
| 6 | 2 |  | 2 |  |
| 7 |  | 1 |  | 1 |
| 8 | 2 |  | 3 |  |
| 9 |  | 2 | 2 |  |
| 10 | 3 |  |  | 3 |
| 11 | 3 |  |  | 2 |
| 12 |  |  | 1 | 1 |
| 13 | 4 |  |  | 4 |
| 14 | 4 | 4 |  |  |

Note that several cells are empty; this means a code was not supplied by a rater. For example, for Passage 1, Rater 4 did not provide a code. With some passages 2 raters provided codes, 3 raters provided codes, or 4 raters provided codes. Notice also that passage 4 has only one rater, so information from that passage cannot be used to calculate level of agreement since all methods for calculating method of agreement requires at least two raters.

This creates problems for Fleiss’s kappa and even makes it difficult to determine how best to calculate percent agreement because some passages will have more raters than others so this creates a problem of weighting percentages.

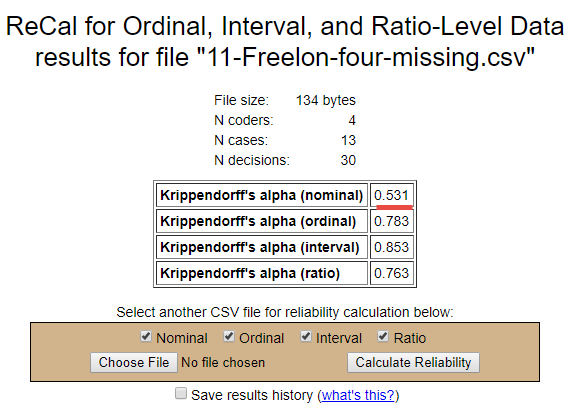
Krippendorff’s alpha, however, is designed to address such missing data and still provide a measure of rater agreement.

**15a. Freelon’s site** [**http://dfreelon.org/utils/recalfront/**](http://dfreelon.org/utils/recalfront/)

To obtain Krippendorff’s alpha with Freelon’s site, replace all missing values with #, then upload the data file as illustrated earlier.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 1 | # |
| 1 | 2 | # | # |
| # | 1 | 1 | 1 |
| 1 | # | # | # |
| 1 | 1 | 2 | 1 |
| 2 | # | 2 | # |
| # | 1 | # | 1 |
| 2 | # | 3 | # |
| # | 2 | 2 | # |
| 3 | # | # | 3 |
| 3 | # | # | 2 |
| # | # | 1 | 1 |
| 4 | # | # | 4 |
| 4 | 4 | # | # |

Results from Freelon’s site; K alpha = .531.



**15b. Geertzen’s site** [**https://nlp-ml.io/jg/software/ira/**](https://nlp-ml.io/jg/software/ira/)

**15c. SPSS**

See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

**16. High Agreement Yet Low Kappa and Alpha**

Measures of rater agreement often provide low values when high levels of agreement exist among raters. The table below shows 20 passages coded by four raters using the four coding categories listed below. Note that all raters agree on every passage except for passage 20.

Despite 95.2% agreement, the other measures of agreement are below acceptable levels: Fleiss’ kappa = .316, mean Cohen’s kappa = .244, and Krippendorff’s alpha = .325.

1 = Positive statement

2 = Negative statement

3 = Neutral statement

4 = Other unrelated statement/Not applicable

The problem with these data is lack of variability in codes. When most raters assign one code predominately, then measures of agreement can be misleadingly low, as demonstrated in this example. This is one reason I recommend always reporting percent agreement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Passage | Rater1 | Rater2 | Rater3 | Rater4 |
| 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 |
| 6 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 |
| 8 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | 1 | 1 |
| 10 | 1 | 1 | 1 | 1 |
| 11 | 1 | 1 | 1 | 1 |
| 12 | 1 | 1 | 1 | 1 |
| 13 | 1 | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | 1 |
| 15 | 1 | 1 | 1 | 1 |
| 16 | 1 | 1 | 1 | 1 |
| 17 | 1 | 1 | 1 | 1 |
| 18 | 1 | 1 | 1 | 1 |
| 19 | 1 | 1 | 1 | 1 |
| 20 | 4 | 3 | 2 | 1 |

Results from Freelon’s site.



**17. Patterns of Response, Bias in Coding Categories, Kappa Paradoxes**

See the link below for details:

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/11-coder-agreement/11-Griffin-agreement-nominal-codes-spr2018.pdf>

**References**

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20:37-46, 1960.

Fleiss (1971) Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76, 378-382.

<http://www.bwgriffin.com/gsu/courses/edur9131/content/Fleiss_kappa_1971.pdf>

Hayes, A. F. & Krippendorfff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77-89.

<http://www.bwgriffin.com/gsu/courses/edur9131/content/coding_reliability_2007.pdf>

Hallgren, K.A. (2012). Computing Inter-rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant. Psychol., 23-34.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/pdf/nihms372951.pdf>

Hruschka, D.J., Schwartz, D., St. John, D.C., Picone-Decaro, E., Jenkins, R.A., & Carey, J.W. (2004). Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. Field Methods, 16, 307-331.

Joyce (2013) Blog Entry: Picking the Best Intercoder Reliability Statistic for Your Digital Activism Content Analysis

[http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activismcontent-analysis](http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis)

Krippendorfff, K. (2011). Agreement and Information in the Reliability of Coding. Communication Methods and Measures, 5 (2), 93-112. <http://dx.doi.org/10.1080/19312458.2011.568376>

<http://repository.upenn.edu/cgi/viewcontent.cgi?article=1286&context=asc_papers>

Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding." Public Opinion Quarterly, 19(3), 321-325.

<http://en.wikipedia.org/wiki/Scott's_Pi>

Warrens, M. J. (2010). Inequalities between multi-rater kappas. Adv Data Anal Classif, 4, 271-286.

[https://openaccess.leidenuniv.nl/bitstream/handle/1887/16237/Warrens\_2010\_ADAC\_4\_271\_286.pdf?sequence=](https://openaccess.leidenuniv.nl/bitstream/handle/1887/16237/Warrens_2010_ADAC_4_271_286.pdf?sequence=2)

[2](https://openaccess.leidenuniv.nl/bitstream/handle/1887/16237/Warrens_2010_ADAC_4_271_286.pdf?sequence=2)

Viera, A.J. & Garrett, J.M. (2005). Understanding interobserver agreement: The Kappa statistic. Family Medicine. 360363.

<http://www.bwgriffin.com/gsu/courses/edur9131/content/Kappa_statisitc_paper.pdf>