07a Coder Agreement for Nominal/Categorical Data

Topics

1. Why Assess Agreement among Coders?

When developing codes and coding responses to open-ended items, it is important that researchers and consumers of the researchers' reports have faith that the data evaluated are credibly reflected in codes, categories, and themes developed by the researcher.

A good practice to follow is to use multiple coders for each unit of text received in questionnaire responses. This means coders who were not part of code/category/theme development process should be trained to understand what each code and category means. They should also be trained how to read questionnaire response text and assign codes from a code sheet or codebook. At least two coders, more than two is preferable, should read and code the same text, and their coding should be evaluated to learn whether they interpreted the questionnaire response text and assigned code similarly. To the extent that coders arrive at different conclusions, this raises questions about credibility of data interpretation.

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. Scale of Coded Data

Recall scales of measurement which include nominal, ordinal, interval, and ratio. Codes used to classify text responses to open-ended items may fit in any of these four scales, although nominal is the most common, ordinal a distant second, and interval/ratio are rare. It is important to identify the scale of coded data from open-ended items because scale determines, in part, which measure is used to assessing inter-rater agreement among coders.

As noted, many of the types of responses received to open-ended questionnaire items result in coded data that forms a nominal scale. For example, Moore and Griffin (2006) asked authors of published studies that appeared in several education-related journals what they perceived to be benefits of co-authoring publications as compared to single-authored work. Responses were coded and presented in Figure 1. The first category is "Quality of Work/Ideas" and it consists of five codes:

1. Diversity of Perspective in work/Ideas
2. Clearer Thinking/Stronger Presentation/Better Written Work
3. Coauthor Peer Review of Work/Ideas
4. Other Quality of Work/Ideas
5. Synthesis of Ideas

Since there is no inherent rank to these codes, the data represented by these codes are nominal in scale.

Figure 1: Moore and Griffin's (2006) Table 2: Perceived Benefits of Coauthored Publications

	Percentage of Respondents [1]	Number of Times Category Referenced [2]
Quality of Work/Ideas	65.0 (39)	
Diversity of Perspective in Work/Ideas		20
Clearer Thinking/Stronger Presentation/Better Written Work		17
Coauthor Peer Review of Work/Ideas		9
Other Quality of Work/Ideas		4
Synthesis of Ideas		3
Division of Labor/Workload	41.7 (25)	
Synthesis of Specialist Skills/Complementary Contributions of Authors		16
Shared Responsibility		2
Other Division of Labor/Workload		9
Collaboration	38.3 (23)	
Sharing of Ideas		8
Builds Community among Academics/ Interaction Among Colleagues		5
Emotional Support		4
General Enjoyment of Collaboration		3
Enables More Extensive Research		2
Motivation to Complete Task		2
Other Collaboration		5
Professional Development	30.0 (18)	
Mentor Novice Writers		9
Learn from Experienced Professionals		5
Enhanced Vita with Less Work		4
Other Professional Development		2

Note: The "Other" category of responses represents responses that could be classified into a given main grouping (such as Professional Development, Collaboration, etc.), but could not be determined to fit within one of the sub-categories for that grouping.

[1] Numbers in parentheses indicate the number of respondents out of 60 who provided a response that fit within a main grouping, e.g., 18 respondents indicated that some aspect of "Professional Development" was used to determine coauthorship.

[2] This column is a simple count of the number of times a specific reason was given for recognition of coauthorship. This column may sum to more than 60 since multiple reasons were often listed by each respondent.

While responses to open-ended items are rarely coded with ordinal type scales, it is common in education for some achievement test responses to be evaluated using an ordinal scale. The Scholastic Aptitude Test (SAT), for example, uses a 1 to 4 scoring rubric for grading essays, which are text responses to questionnaire items.

Below, in Figure 2, the College Board, author of the SAT, explains that essays are scored on three dimensions: reading, analysis, and writing. Each essay is evaluated by two raters, with each dimension receiving a score of 1 to 4 from each rater, for a total score of 2 to 8 per dimension.

Figure 2: SAT Essay Scoring

How the SAT Essay Is Scored

Responses to the optional SAT Essay are scored using a carefully designed process.

- Two different people will read and score your essay.
- Each scorer awards 1–4 points for each dimension: reading, analysis, and writing.
- The two scores for each dimension are added.
- You'll receive three scores for the SAT Essay—one for each dimension—ranging from 2–8 points.
- There is no composite SAT Essay score (the three scores are not added together) and there are no percentiles.

We train every scorer to hold every student to the same standards, the ones shown on this page.

Source: https://collegereadiness.collegeboard.org/sat/scores/understanding-scores/essay

The College Board provides some details of their scoring rubric, but the essence is given in the summary below. This summary represents partial scoring criteria for Reading dimension of essays.

Score = 1: Demonstrates little to no comprehension of the source text.
Score = 2: Demonstrates some comprehension of the source text.
Score = 3: Demonstrates effective comprehension of the source text.
Score = 4: Demonstrates thorough comprehension of the source text.

This rubric shows that scores increase as demonstrated understanding of the material increases. Given this, SAT's scoring plan produces ordinal-level data, although some may treat these data as interval for analysis purposes.

In summary, to use the inter-rater (or inter-coder) agreement measures described below, it is important to identify the measurement scale for coded data.

3. 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

3a. 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 a 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 = $\frac{12}{18} \times 100$ = .6667 × 100 = 66.67%

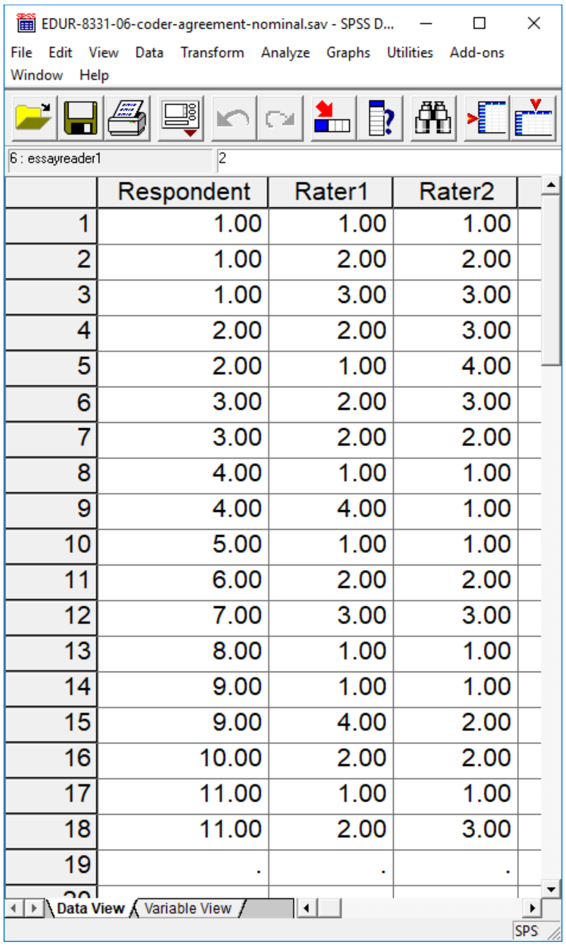
(b) Percent Agreement for Two Raters, SPSS

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/edur8331/edur8331-presentations/EDUR-8331-06-coder-agreement-nominal.sav

For these data rater 1 is labeled **Rater1** and rater 2 is labeled **Rater2**. For now, ignore other data found in the SPSS file.



	Respondent	Rater1	Rater2
1	1.00	1.00	1.00
2	1.00	2.00	2.00
3	1.00	3.00	3.00
4	2.00	2.00	3.00
5	2.00	1.00	4.00
6	3.00	2.00	3.00
7	3.00	2.00	2.00
8	4.00	1.00	1.00
9	4.00	4.00	1.00
10	5.00	1.00	1.00
11	6.00	2.00	2.00
12	7.00	3.00	3.00
13	8.00	1.00	1.00
14	9.00	1.00	1.00
15	9.00	4.00	2.00
16	10.00	2.00	2.00
17	11.00	1.00	1.00
18	11.00	2.00	3.00
19	.	.	.

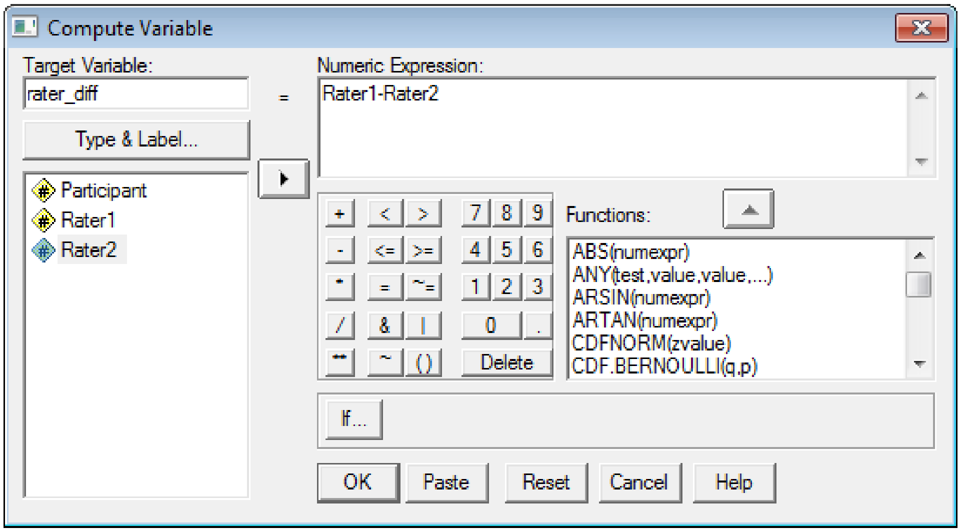
(2) Calculate difference of reviewer scores

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**



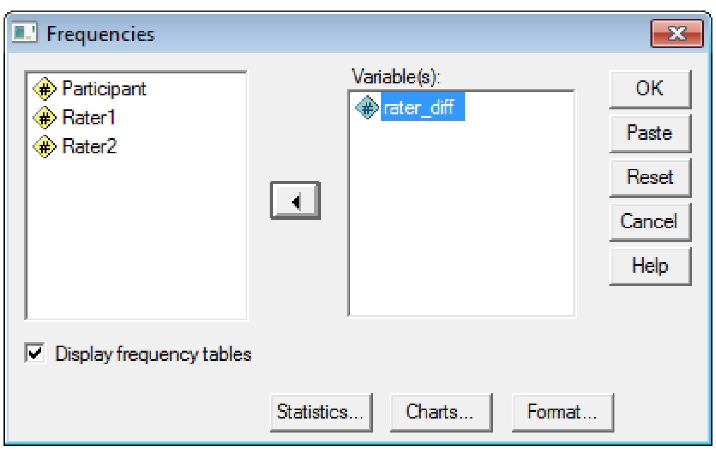
Click "OK" to run the compute command.

(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	5.6	5.6	5.6
	-1.00	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.

r2r3diff

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-2.00	1	2.8	5.6	5.6
	-1.00	2	5.6	11.1	16.7
	.00	13	36.1	72.2	88.9
	2.00	1	2.8	5.6	94.4
	3.00	1	2.8	5.6	100.0
	Total	18	50.0	100.0	
Missing	System	18	50.0		
Total		36	100.0		

3b. 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%	

(Instructor's note to self: The calculations of average percentage agreement shown above match the formula provided by Fleiss (1971; see page 379 for average agreement formula)).

3c. 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			
		1.00	2.00	3.00	Total
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 consider random agreement opportunities.

3d. Measures of Agreement among Two Raters other than Percentage Agreement

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. The material that follows presents alternative measures of rater agreement that adjust for possible random agreement among raters.

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.

Interpretation of Krippendorff's alpha:

> When human lives hang on the results of a content analysis, whether they inform a legal decision or tip the scale from peace to war, decision criteria have to be set far higher than when a content analysis is intended to merely support scholarly arguments. In case of the latter, to be sure that the data under consideration are at least similarly interpretable by other scholars (as represented by different coders), I suggested elsewhere to require $\alpha \geq .800$, and where tentative conclusions are still acceptable, $\alpha \geq .667$ (Krippendorff, 2004, p. 241).

In summary, for most research purposes a K-alpha of .66 or greater is desired.

3f. 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

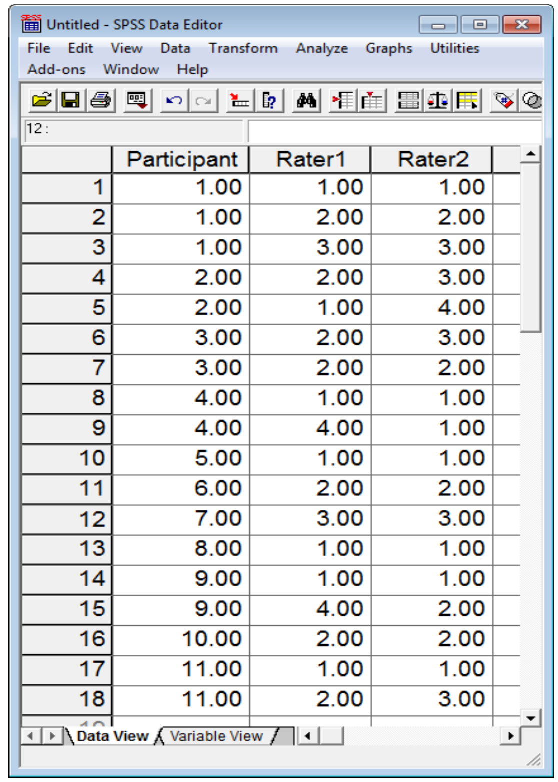
Source: Viera & Garrett, 2005, Understanding interobserver agreement: The Kappa statistic. Family Medicine.

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

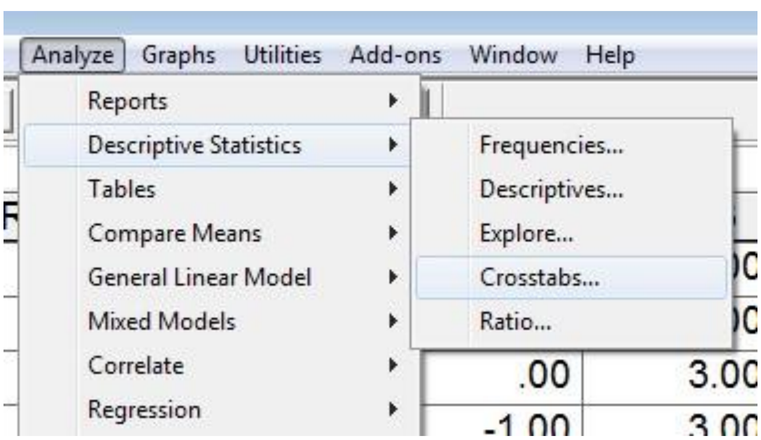
Codes from each rater must be linked or matched for reliability analysis to work properly. Note these are the same data used to calculate percentage agreement. An example of data entry in SPSS is also provided.

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



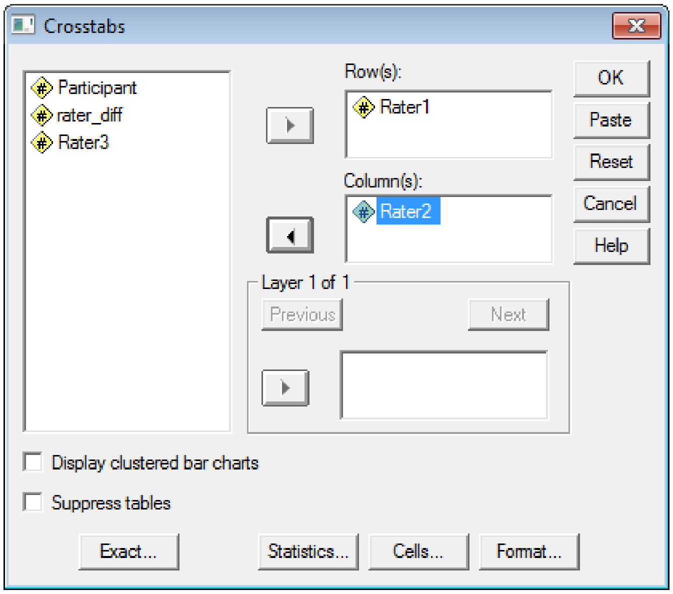
To run kappa, use crosstabs command:

Analyze → Descriptive Statistics → Crosstabs

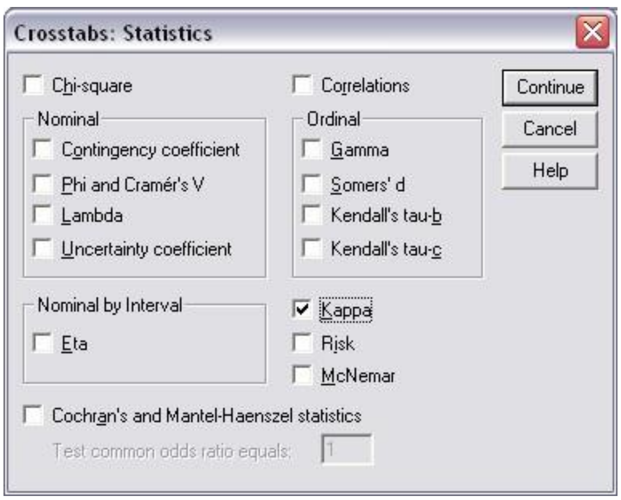


With the Crosstabs pop-up menu, move the raters' coding to the Row and Column boxes. One rater should be identified as the row, the other as the column – which rater is assigned to row or column is not important.

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.

(b) Cohen's Kappa via SPSS: Weighted Cases

Sometimes the number of data points generated can be very large. In such cases the pattern of codes may be entered into SPSS to help reduce the data entry burden. In other cases only a summary table of results is provided. It may look like this, for example:

Results of Review Ratings

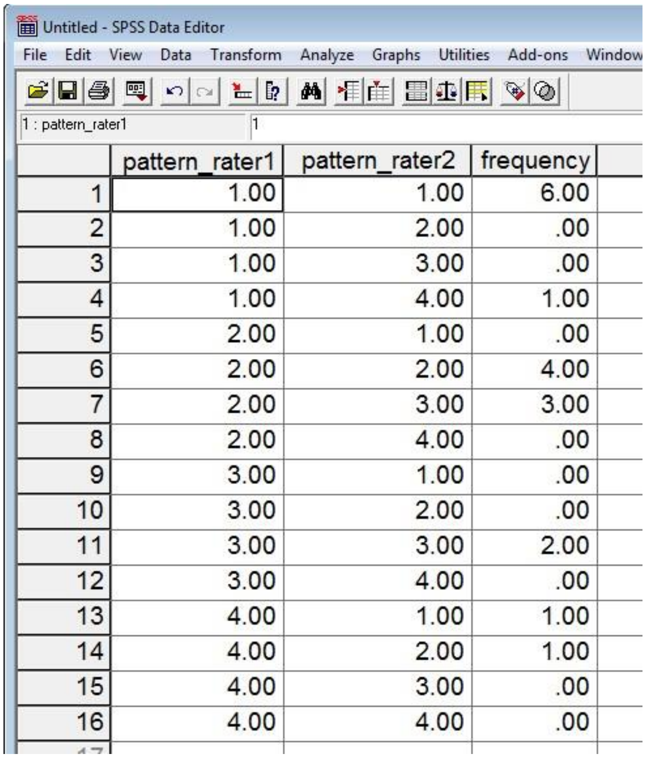
		Rater 2			
		1 = Positive	2 = Negative	3 = Neutral	4 = Other
Rater 1	1 = Positive	**6**	0	0	1
	2 = Negative	0	**4**	3	0
	3 = Neutral	0	0	**2**	0
	4 = Other	1	1	0	**0**

Note: Numbers indicate counts, e.g., there are 6 cases in which raters 1 and 2 agreed the statement was positive.

It is useful to record all response pattern options first, and then count those that occur. This includes those patterns that are not found among the reviewers. See below for examples which frequency of pattern = 0.

Original Ratings			Pattern of Ratings and Frequency of Pattern		
Reviewer 1	Reviewer 2		Pattern Reviewer 1	Pattern Reviewer 2	Frequency of Pattern
1	1		1	1	6
2	2		1	2	0
3	3		1	3	0
2	3		1	4	1
1	4		2	1	0
2	3		2	2	4
2	2		2	3	3
1	1		2	4	0
4	1		3	1	0
1	1		3	2	0
2	2		3	3	2
3	3		3	4	0
1	1		4	1	1
1	1		4	2	1
4	2		4	3	0
2	2		4	4	0
1	1				
2	3				

Example of data entry in SPSS appears below.

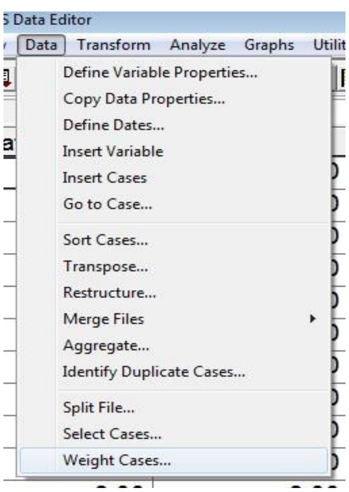


Untitled - SPSS Data Editor

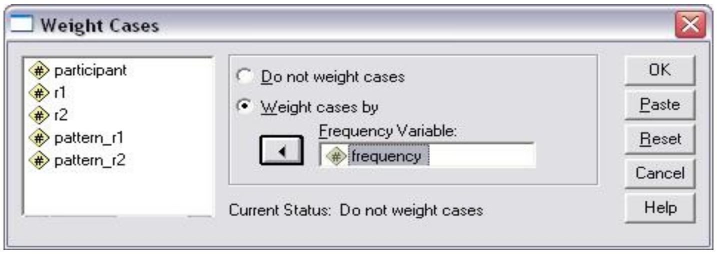
	pattern_rater1	pattern_rater2	frequency
1	1.00	1.00	6.00
2	1.00	2.00	.00
3	1.00	3.00	.00
4	1.00	4.00	1.00
5	2.00	1.00	.00
6	2.00	2.00	4.00
7	2.00	3.00	3.00
8	2.00	4.00	.00
9	3.00	1.00	.00
10	3.00	2.00	.00
11	3.00	3.00	2.00
12	3.00	4.00	.00
13	4.00	1.00	1.00
14	4.00	2.00	1.00
15	4.00	3.00	.00
16	4.00	4.00	.00

When patterns of coding are entered into SPSS, one must inform SPSS about the weighting of each pattern – the frequency of each pattern. To correctly weight cases, use the Weight Cases command:

Data→ Weight Cases



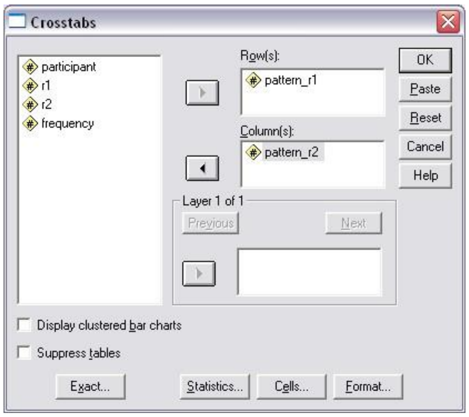
Once the pop-up window appears, place a mark next to "Weight cases by," select the weight variable (in this example it is "frequency"), move that variable to the "Frequency Variable" box. Click on the "OK" button to finish assigning variable weights. This process is illustrated in the image below.



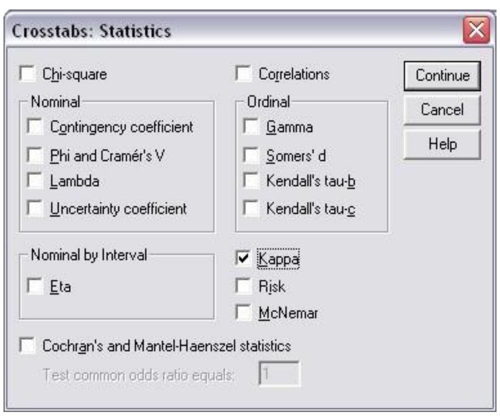
Once the weighting variable is identified, one may now run the crosstabs command as illustrated earlier:

Analyze → Descriptive Statistics → Crosstabs

With the Crosstabs pop-up menu, move the raters' **pattern** coding to the Row and Column boxes. One rater's **pattern** should be identified as the row, the other as the column – which raters' **pattern** is assigned to row or column is not important. This is illustrated in the image below.



Next, select "Statistics" then place mark next to "Kappa", click "Continue" then "OK" to run the analysis.



In this case kappa is, again, .526.

(c) SPSS Limitation with Cohen's kappa

Update: Newer versions of SPSS (at least version 21, maybe earlier editions too) do not suffer from the problem described below.

SPSS cannot calculate kappa if one rater does not use all of the same rating categories as another rater. Suppose two raters are asked to rate an essay as either:

1 = pass
2 = pass with revisions
3 = fail

Their ratings appear in the table below. Note that Rater 1 uses the three categories of 1, 2, and 3, but Rater 2 does not assign a rating of 3 to any essay.

Essay	Essay Rater 1	Essay Rater 2
1	1	1
2	1	1
3	1	1
4	2	2
5	2	2
6	2	2
7	2	2
8	2	2
9	2	2
10	2	2
11	3	2
12	3	2
13	3	2
14	3	2

UCLA Statistical Consulting Group provided a workaround explained here in the link below, but that link is now defunct.

http://www.ats.ucla.edu/stat/spss/faq/kappa.htm

I provide an image of their explanation in Figure 3 below.

Figure 3: UCLA Statistics Consulting Group SPSS Cohen Kappa Solution to Unequal Categories by Raters

SPSS FAQ

How can I calculate a kappa statistic for variables with unequal score ranges?

Suppose we would like to compare two raters using a kappa statistic but the raters have different range of scores. This situation most often presents itself where one of the raters did not use the same range of scores as the other rater.

Let us consider an example where two graduate students where asked to rate 12 movies based on a scale from 1-3. One rater used all of the three scores possible while rating the movies whereas the other student did not like any of the movies and therefore rated all of them as either a 1 or a 2. Thus, the range of scores is the not the same for the two raters.

To obtain the kappa statistic in SPSS we are going to use the **crosstabs** command with the **statistics = kappa** option. By default, SPSS will only compute the kappa statistics if the two variables have exactly the same categories, which is not the case in this particular instance. We can get around this problem by adding a fake observation and a weight variable shown below. The weight variable takes value of 1 for all the real observations and value of 0.00001 (something very small) for the fake observation that we have just added. The trick is then to weight the observations using the **weight** command.

```
data list list
/rater1 rater2.
begin data.
  1        1
  1        1
  1        1
  1        1
  2        2
  2        2
  2        2
  2        2
  3        2
  3        2
  3        2
  3        2
end data.

save outfile = kappa.
data list list
/rater1 rater2.
begin data.
  3        3
end data.

add files file = *
/file = kappa.
exe.

compute weight = 1.
if ( rater1 =3 & rater2 =3 ) weight = .00001.
exe.

weight by weight.
crosstabs
  /tables=rater1 by rater2
  /statistics=kappa.
```

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T (b)	Approx. Sig.
Measure of Agreement	Kappa	.500	.156	3.000	.003
N of Valid Cases		12			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

UCLA's solution requires using weighted data rather than unweight (ungrouped) data. Find the pattern of responses as explained earlier:

Essay	Essay Rater 1	Essay Rater 2		Pattern Rater 1	Pattern Rater 2	Frequency of Pattern
1	1	1		1	1	3
2	1	1		1	2	0
3	1	1		1	3	0
4	2	2		2	1	0
5	2	2		2	2	7
6	2	2		2	3	0
7	2	2		3	1	0
8	2	2		3	2	4
9	2	2		3	3	0
10	2	2				
11	3	2				
12	3	2				
13	3	2				
14	3	2				

For rater 2 there are no values of 3 used for rating essays; as the pattern of ratings above show, the frequency of rater 2 assigning a value of 3 is 0 (see highlighted cells). To fool SPSS into calculating kappa, replace any one of the 0 frequencies highlighted above with a very small value, such as .0001. Use a small number so it does not influence calculation of kappa. See below:

Essay	Essay Rater 1	Essay Rater 2		Pattern Rater 1	Pattern Rater 2	Frequency of Pattern
1	1	1		1	1	3
2	1	1		1	2	0
3	1	1		1	3	0
4	2	2		2	1	0
5	2	2		2	2	7
6	2	2		2	3	0
7	2	2		3	1	0
8	2	2		3	2	4
9	2	2		3	3	.0001
10	2	2				
11	3	2				
12	3	2				
13	3	2				
14	3	2				

Now execute the crosstabs command again with these data (remember to assign Data-> Weight Case) and SPSS should provide the following kappa results.

b1p * b2p Crosstabulation

		b2p			Total
		1.00	2.00	3.00	
b1p	1.00	3	0	0	3
	2.00	0	7	0	7
	3.00	0	4	0	4
Total		3	11	0	14

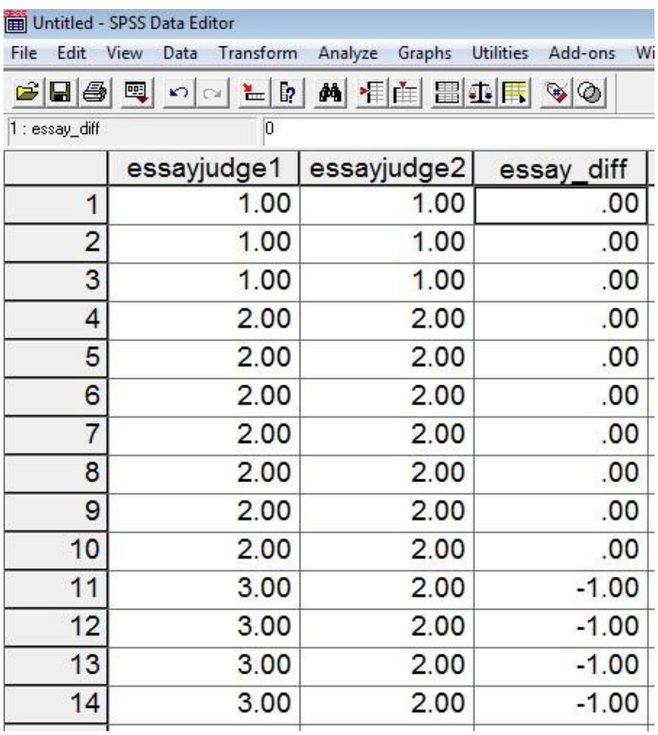
Symmetric Measures

		Value	Asymp.Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement	Kappa	.491	.177	3.159	.002
N of Valid Cases		14			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Although you cannot see it, the frequency cell highlighted in gold above actually has the value .0001 but has been rounded to 0 by SPSS. Cohen's kappa is .491 for these data.

The percentage agreement for these data can be found as noted earlier by calculating the difference between judges then finding the percentage of agreements. The SPSS file with differences calculated follows:



	essayjudge1	essayjudge2	essay_diff
1	1.00	1.00	.00
2	1.00	1.00	.00
3	1.00	1.00	.00
4	2.00	2.00	.00
5	2.00	2.00	.00
6	2.00	2.00	.00
7	2.00	2.00	.00
8	2.00	2.00	.00
9	2.00	2.00	.00
10	2.00	2.00	.00
11	3.00	2.00	-1.00
12	3.00	2.00	-1.00
13	3.00	2.00	-1.00
14	3.00	2.00	-1.00

The frequency display appears below.

essay_diff

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1.00	4	25.0	28.6	28.6
	.00	10	62.5	71.4	100.0
	Total	14	87.5	100.0	
Missing	System	2	12.5		
Total		16	100.0		

The percentage agreement is 71.4% (again, note that one should always use the "Valid Percent" column since it ignores missing data for calculating category percentages).

3g. 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)

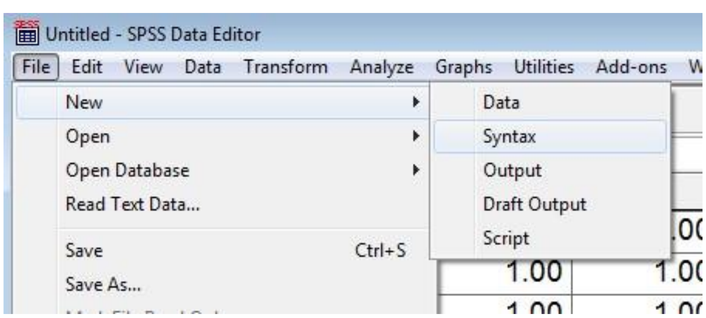
http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Krippendorff-alpha-SPSS.txt

(a) K alpha with SPSS

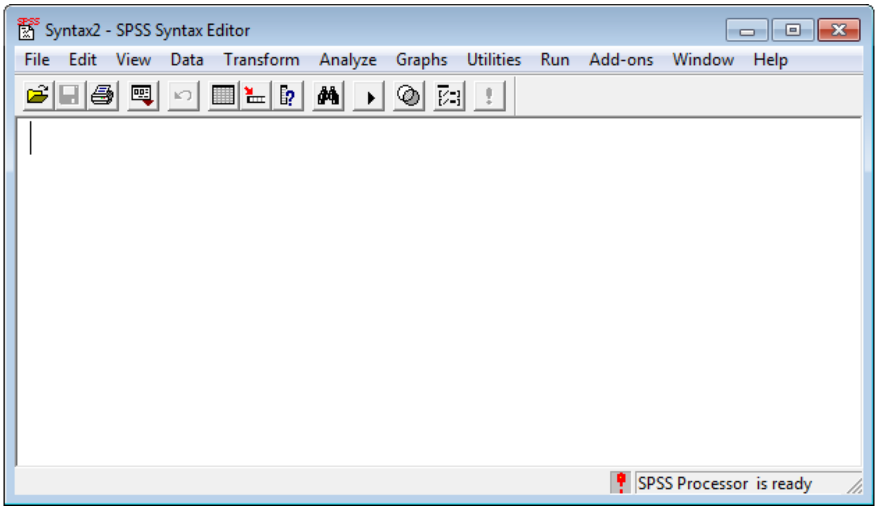
Note – This option does not work well with all versions of SPSS and is more cumbersome than using Freelon's webpage which is explained below in the next section (about four pages down). I recommend skipping directly to Freelon's page to obtain Krippendorff's alpha and others measures of agreement.

To copy and paste the K alpha commands into SPSS, do the following:

File → New → Syntax



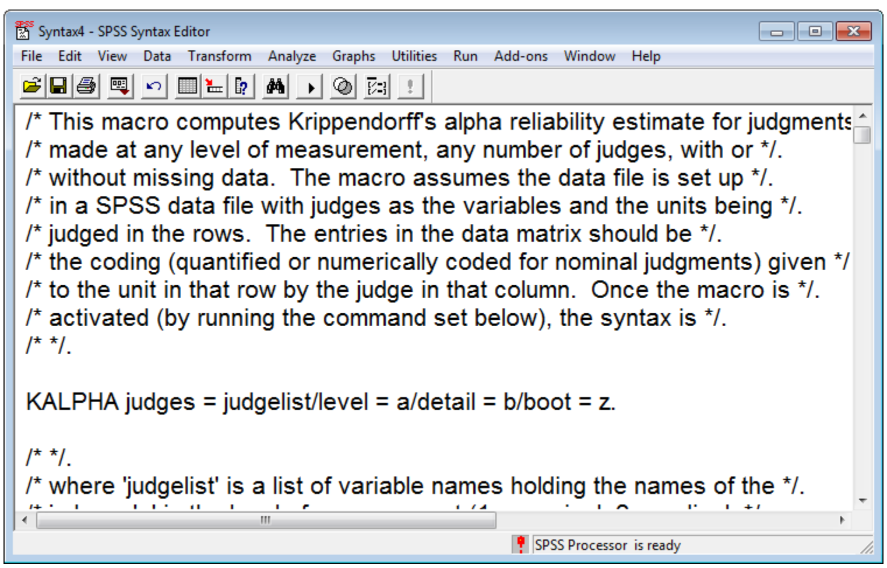
This opens a syntax window that should be similar to this window:



Now open the K alpha commands from this link

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Krippendorff-alpha-SPSS.txt

Next, copy and paste everything find at that link into the SPSS syntax window. When you finish, it should look like this:



To make this syntax work, four bits of the command line must be changed. The command line is the isolated line above that reads:

KALPHA judges = judgelist/level = a/detail = b/boot = z.

judges = judgelist

These are the raters which form columns in SPSS

level = a

This is the scale of measurement of ratings with

1 = nominal

2 = ordinal
3 = interval
4 = ratio

Since we are dealing with ratings that are nominal, select 1 here.

detail = b

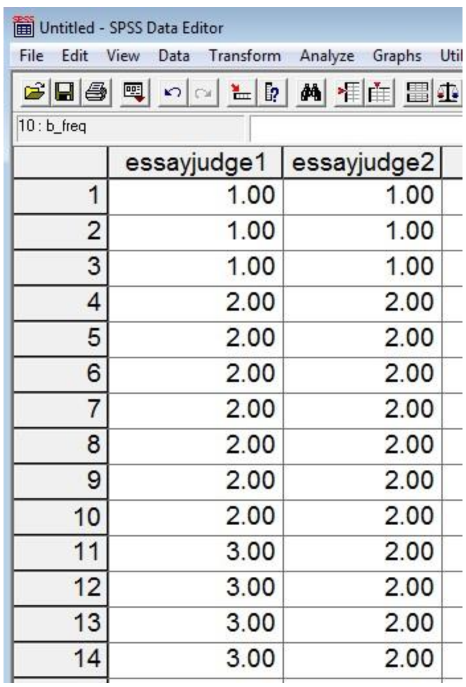
Specify 0 or 1 here; by default select 1 to see calculations.

boot = z

This option allows one to obtain bootstrapped standard errors for the K alpha estimate. For our purposes we won't request standard errors so place 0 for this option. If you wanted standard errors, the minimum replications would be 1000.

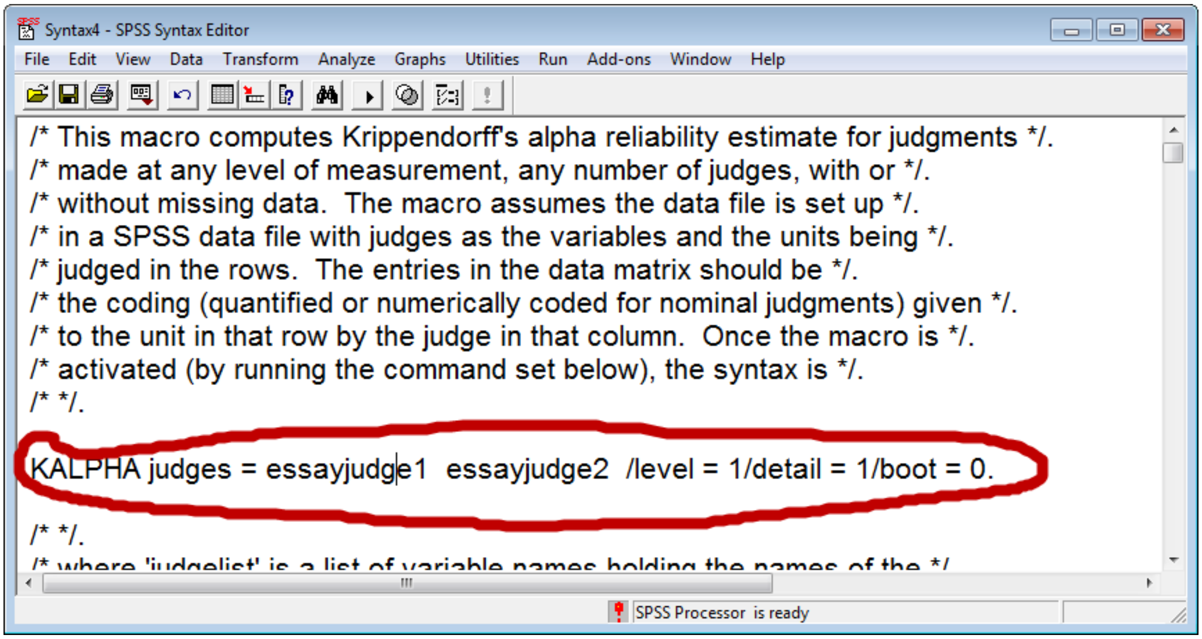
To obtain K alpha for the essay data below, make the following changes to the Kalpha command in the syntax window:

KALPHA judges = essayreader1 essayreader2 /level = 1/detail = 1/boot = 0.



	essayjudge1	essayjudge2
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	2.00	2.00
5	2.00	2.00
6	2.00	2.00
7	2.00	2.00
8	2.00	2.00
9	2.00	2.00
10	2.00	2.00
11	3.00	2.00
12	3.00	2.00
13	3.00	2.00
14	3.00	2.00

The SPSS syntax window now looks like this:



To run this command, place the mouse cursor within the KALPHA command (anywhere in the command sentence), and then click on "Run Current" button which looks like this  on my version of SPSS.

K Alpha SPSS output

Matrix

```
Run MATRIX procedure:

Krippendorff's Alpha Reliability Estimate

              Alpha     Units    Obsrvrs     Pairs
Nominal       .4706   14.0000     2.0000   14.0000

Judges used in these computations:
 essayjud essayj_1

====================================================

Observed Coincidence Matrix
      6.00       .00       .00
       .00     14.00      4.00
       .00      4.00       .00

Expected Coincidence Matrix
      1.11      4.00       .89
      4.00     11.33      2.67
       .89      2.67       .44

Delta Matrix
       .00      1.00      1.00
      1.00       .00      1.00
      1.00      1.00       .00

Rows and columns correspond to following unit values
      1.00      2.00      3.00

Examine output for SPSS errors and do not interpret if any are found

------ END MATRIX -----
```

Krippendorff argues that values below .80 should be viewed as poor levels of agreement, so this value of .47 suggest problems with rater agreement.

(b) 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/~~
Unfortunately, this site is no longer available; I am retaining the instructions below should the site return.

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)~~

~~Note that Geertzen's site, https://nlp-ml.io/jg/software/ira/, only addresses nominal rating categories. If one has ordinal, interval, or ratio ratings, then calculations from Geertzen's site may be inappropriate.~~

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
```

~~Geertzen's site requires similar data structure, but no commas and each column should have a header identifying the rater. There should be a blank space or tab between ratings and headers, like this:~~

~~rater1 rater2~~
~~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 two files suitable for use with Freelon's and Geertzen's sites.

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Freelon-essay-data.csv

~~http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Geertzen-essay-data.txt~~

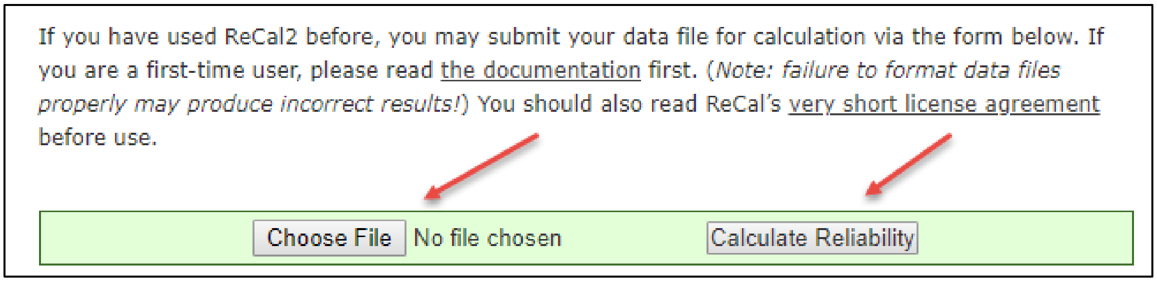
Download both files to your computer, then upload both to the respective websites.

Freelon's site (http://dfreelon.org/utils/recalfront/)

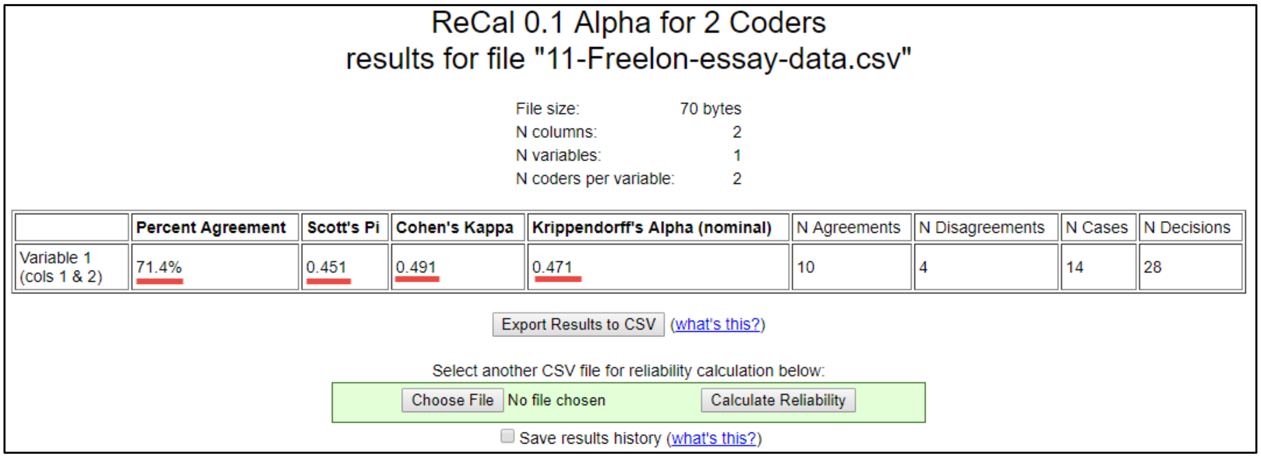
(a) Select the link for ReCal2 for nominal data and 2 coders.

Level of measurement	N of coders	Missing data allowed?	Use
Nominal	2 coders only	No	ReCal2 (includes percent agreement, Scott's pi, Cohen's kappa, and nominal Krippendorff's alpha)
Nominal	2 or more coders	No	ReCal3 (includes pairwise percent agreement, Fleiss' kappa, pairwise Cohen's kappa, and nominal Krippendorff's alpha)
Nominal, ordinal, interval, or ratio	Any N of coders	Yes	ReCal OIR (includes nominal, ordinal, interval, and ratio Krippendorff's alpha with support for missing data)

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



(c) Note results


ReCal 0.1 Alpha for 2 Coders
results for file "11-Freelon-essay-data.csv"

File size: 70 bytes
N columns: 2
N variables: 1
N coders per variable: 2

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (nominal)	N Agreements	N Disagreements	N Cases	N Decisions
Variable 1 (cols 1 & 2)	71.4%	0.451	0.491	0.471	10	4	14	28

Export Results to CSV (what's this?)

Select another CSV file for reliability calculation below:

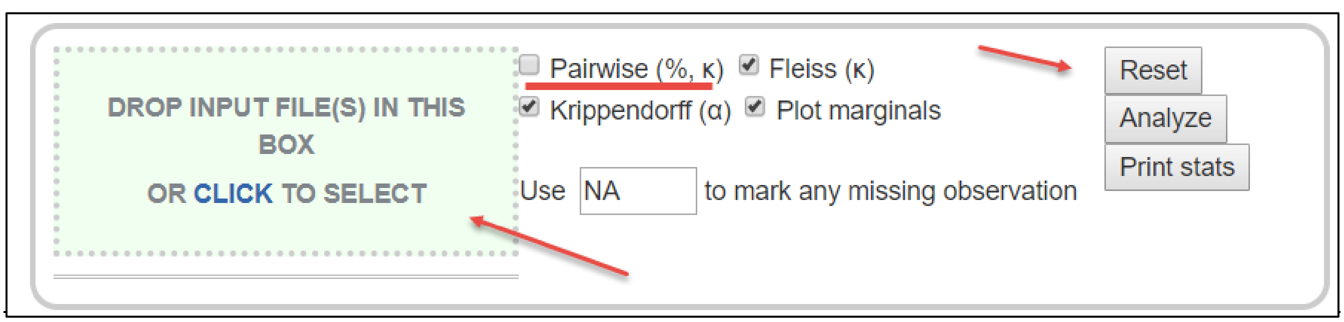
Choose File No file chosen Calculate Reliability

Save results history (what's this?)

Percent agreement = 71.4
Scott's pi = .451
Cohen's kappa = .491
K alpha = .471

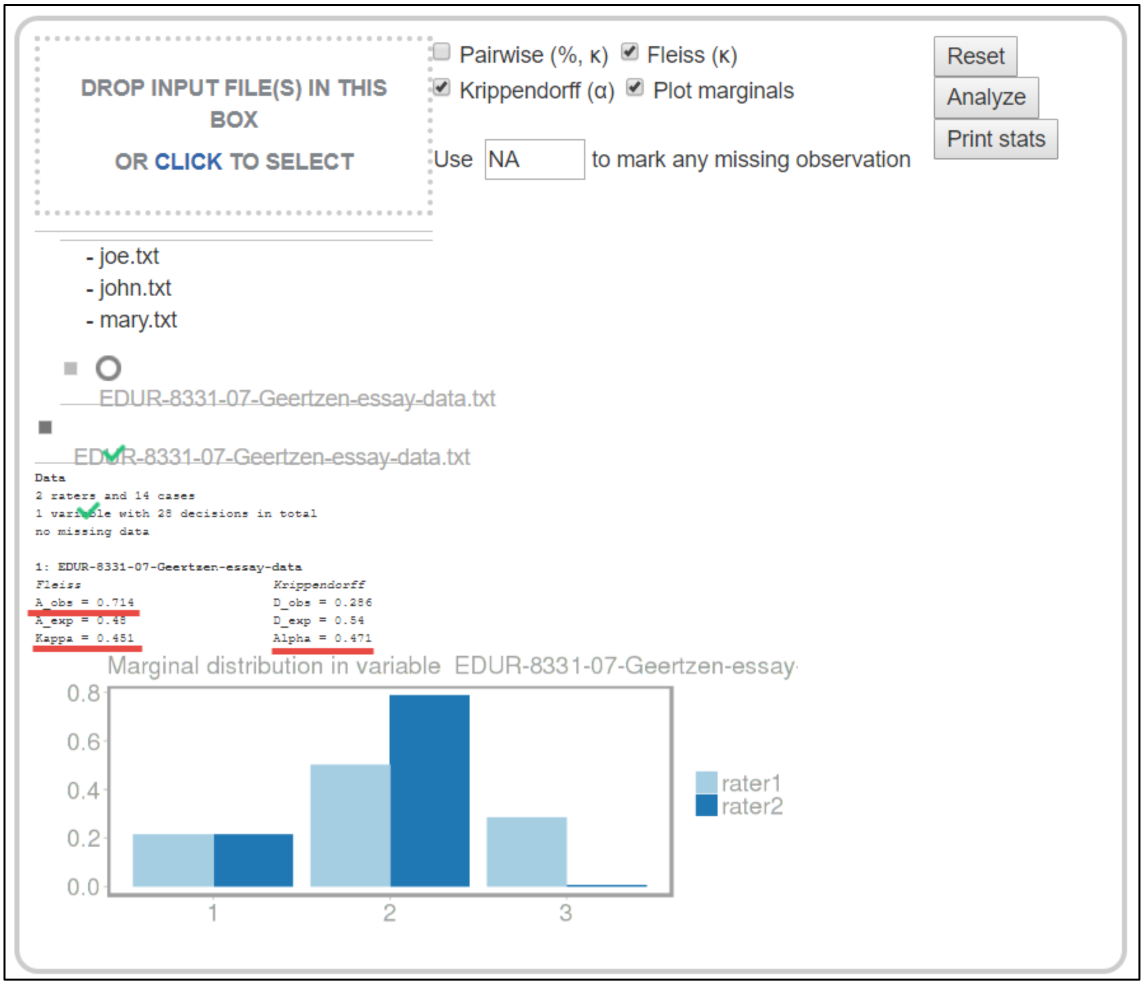
~~**Geertzen's site (https://nlp-ml.io/jg/software/ira/)**~~

~~(a) Click "Reset" then drag the file to the drop box or "Click" to select files from your computer. Unfortunately, I was unable to obtain results with a check next to "Pairwise (%, κ)" so live that box blank otherwise an error will result.~~



~~(b) Once uploaded, click select all options (except for the Pairwise box), then click "Analyze"~~

~~(c) Note output~~



~~Fleiss kappa (Scott's pi) = .451~~
~~K alpha = .471~~
~~Percent agreement = .714 or 71.4%~~

4. Two-coder Supplemental Examples

Both examples display raw data – counts of agreement and disagreement between two raters – in cross-tabulation tables. Below in example 4b I explain how to convert these data into a spreadsheet for analysis of agreement.

4a. Usefulness of Noon Lectures

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

Usefulness of Noon Lectures

		Resident 1— Lectures Helpful?		
		Yes	*No*	*Total*
Resident 2— Lectures Helpful?	*Yes*	**15**	5	20
	No	10	**70**	80
	Total	25	75	100

Answers
Percent agreement = 85.0
Scott's pi = .570
Cohen's kappa = .571
K alpha = .572

Since K alpha is less than .66, one would just this agreement to be less than acceptable.

Data in CSV format (I used Freelon's site and uploaded these data)
http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Supplemental-Example-1.csv

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
Rater 1	Anger	6	0	1	2
	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.

Below is an explanation how to convert this table into a spreadsheet format that can be used by Freelon's site or SPSS to calculate agreement.

The table above contains counts of agreement and disagreement. For example, there are five times both Rater 1 and Rater 2 agreed that the face reviewed displayed Disgust. There are two times Raters 1 thought the face showed Contempt while Rater 2 disagree and thought the face showed Disgust.

The first step in converting the tabled data into a spreadsheet format is to assign numbers to the rating categories. I will use the following:

1 = Anger
2 = Fear
3 = Disgust
4 = Contempt

Now create a spreadsheet type table to expand counts in the table above. For example, Rater 1 and Rater 2 provided 6 ratings that agreed the face showed Anger, so their ratings in numeric form is shown below.

Rater 1	Rater 2	Rater 1 Actual Assessment	Rater 2 Actual Assessment	Combination Count
1	1	Anger	Anger	1
1	1	Anger	Anger	2
1	1	Anger	Anger	3
1	1	Anger	Anger	4
1	1	Anger	Anger	5
1	1	Anger	Anger	6

There was one occurrence where Rater 1 judged the face to show Anger (score of 1) while Rater 2 judged it to show Disgust (score of 3)

Rater 1	Rater 2	Rater 1 Actual Assessment	Rater 2 Actual Assessment	Combination Count
1	3	Anger	Disgust	1

There were two occurrences where Rater 1 judged the face to show Anger (score of 1) while Rater 2 judged it to show Contempt (score of 4).

Rater 1	Rater 2	Rater 1 Actual Assessment	Rater 2 Actual Assessment	Combination Count
1	4	Anger	Contempt	1
1	4	Anger	Contempt	2

This process must be done for each combination in which there is a count greater than 0 in the data table above. The complete spreadsheet conversion is shown below.

Rater 1	Rater 2	Rater 1 Actual Assessment	Rater 2 Actual Assessment	Combination Count
1	1	Anger	Anger	1
1	1	Anger	Anger	2
1	1	Anger	Anger	3
1	1	Anger	Anger	4
1	1	Anger	Anger	5
1	1	Anger	Anger	6
1	3	Anger	Disgust	1
1	4	Anger	Contempt	1
2	2	Fear	Fear	1
2	2	Fear	Fear	2
2	2	Fear	Fear	3
2	2	Fear	Fear	4
2	3	Fear	Disgust	1
2	3	Fear	Disgust	2
3	1	Disgust	Anger	1

3	1	Disgust	Anger	2
3	2	Disgust	Fear	1
3	3	Disgust	Disgust	1
3	3	Disgust	Disgust	2
3	3	Disgust	Disgust	3
3	3	Disgust	Disgust	4
3	3	Disgust	Disgust	5
3	4	Disgust	Contempt	1
4	1	Contempt	Anger	1
4	2	Contempt	Fear	1
4	3	Contempt	Disgust	1
4	3	Contempt	Disgust	2
4	4	Contempt	Contempt	1
4	4	Contempt	Contempt	2
4	4	Contempt	Contempt	3
4	4	Contempt	Contempt	4

Converted to a format that works for Freelon's site, one must use only numbers with no headers and not text. This file can be downloaded from the link provided below.

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Supplemental-Example-2.csv

Answers
Percent agreement = 59.4
Scott's pi = .453
Cohen's kappa = .453
K alpha = .462

Since K alpha is less than .66, one would just this agreement to be less than acceptable.

5. 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%	

6. 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

7. 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 using Freelon's site.

8. 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.

9. Three Rater Example: Percent Agreement, Cohen's Kappa Mean, Fleiss' kappa, Krippendorff's alpha

The three-rater data, presented above in "**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.

9a. Freelon's site 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/edur8331/edur8331-presentations/EDUR-8331-07-Freelon-three-raters.csv

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

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

Level of measurement	N of coders	Missing data allowed?	Use
Nominal	2 coders only	No	ReCal2 (includes percent agreement, Scott's pi, Cohen's kappa, and nominal Krippendorff's alpha)
Nominal	2 or more coders	No	ReCal3 (includes pairwise percent agreement, Fleiss' kappa, pairwise Cohen's kappa, and nominal Krippendorff's alpha)
Nominal, ordinal, interval, or ratio	Any N of coders	Yes	ReCal OIR (includes nominal, ordinal, interval, and ratio Krippendorff's alpha with support for missing data)

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

Choose File EDUR-8331-0…-raters.csv Calculate Reliability

Results are reported below

ReCal 0.1 Alpha for 3+ Coders
results for file "11-Freelon-three-raters.csv"

File size:	162 bytes
N coders:	3
N cases:	18
N decisions:	54

Average Pairwise Percent Agreement

Average pairwise percent agr.	Pairwise pct. agr. cols 1 & 3	Pairwise pct. agr. cols 1 & 2	Pairwise pct. agr. cols 2 & 3
66.667%	61.111%	66.667%	72.222%

Fleiss' Kappa

Fleiss' Kappa	Observed Agreement	Expected Agreement
0.518	0.667	0.308

Average Pairwise Cohen's Kappa

Average pairwise CK	Pairwise CK cols 1 & 3	Pairwise CK cols 1 & 2	Pairwise CK cols 2 & 3
0.521	0.435	0.526	0.602

Krippendorff's Alpha (nominal)

Krippendorff's Alpha	N Decisions	$\Sigma_c o_{cc}$***	$\Sigma_c n_c(n_c - 1)$***
0.527	54	36	844

***These figures are drawn from Krippendorff (2007, case C.)

Export Results to CSV (what's this?)

Select another CSV file for reliability calculation below:

Choose File No file chosen Calculate Reliability

Save results history (what's this?)

Percentage agreement = 66.7
Mean Cohen's kappa (pairwise kappa) = .521
Fleiss' kappa = .518
Krippendorff's alpha = .527

All suggest low agreement among raters.

~~**9b. Geertzen's site** https://nlp-ml.io/jg/software/ira/~~

~~The data file for Geertzen's site should follow the format shown below.~~

~~R1~~	~~R2~~	~~R3~~
~~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~~

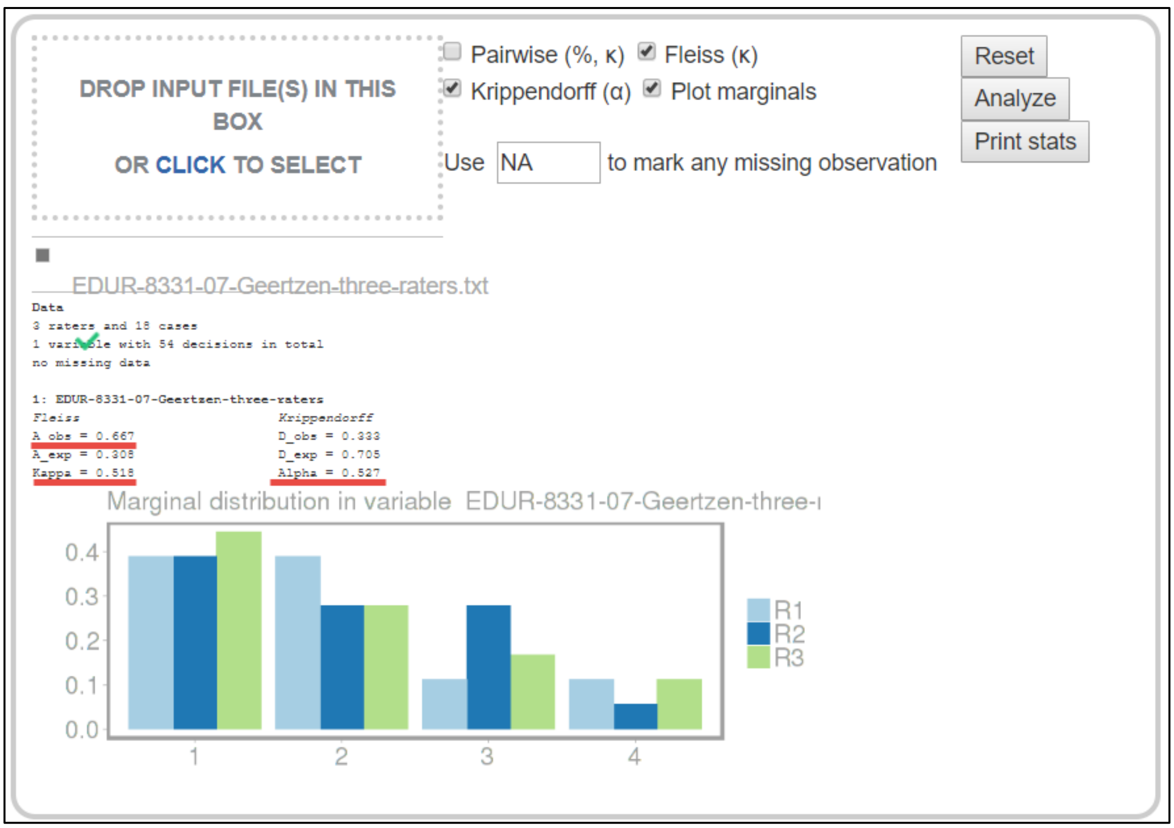
~~3~~	~~3~~	~~3~~
~~1~~	~~1~~	~~1~~
~~1~~	~~1~~	~~2~~
~~4~~	~~2~~	~~2~~
~~2~~	~~2~~	~~2~~
~~1~~	~~1~~	~~1~~
~~2~~	~~3~~	~~4~~

~~Here is a text file with these data:~~

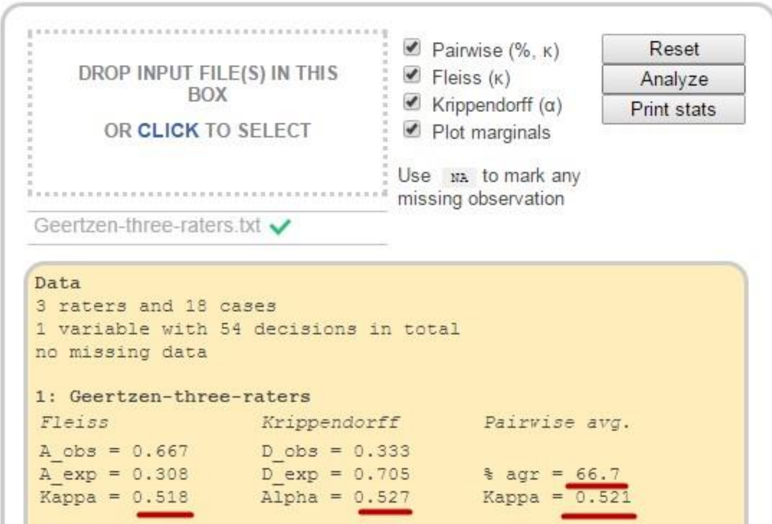
~~http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-07-Geertzen-three-raters.txt~~

~~Follow the steps outlined earlier – (a) Click Reset if any results are current presented, (b) upload or drag the data file to the input box, and (c) select those statistics of interest.~~

~~As noted before, I was unsuccessful in obtaining Cohen's kappa and Pairwise percentages, so remove the check mark from the Pairwise box and the page is then able to estimate Fleiss's kappa and Krippendorff's alpha.~~



~~Below is output from an earlier version of the page with functioning Pairwise percentages and Cohen's kappa.~~



~~Percentage agreement = 66.7 (same as reported in hand calculation)~~
~~Mean Cohen's kappa (pairwise kappa) = .521 (same as found with mean kappa in SPSS)~~
~~Fleiss kappa = .518~~
~~Krippendorff alpha = .527~~

9c. SPSS
The three-rater data noted above are entered into SPSS as follows:

	r1	r2	r3
1	1.00	1.00	1.00
2	2.00	2.00	2.00
3	3.00	3.00	3.00
4	2.00	3.00	3.00
5	1.00	4.00	1.00
6	2.00	3.00	1.00
7	2.00	2.00	4.00
8	1.00	1.00	1.00
9	4.00	1.00	1.00
10	1.00	1.00	1.00
11	2.00	2.00	2.00
12	3.00	3.00	3.00
13	1.00	1.00	1.00
14	1.00	1.00	2.00
15	4.00	2.00	2.00
16	2.00	2.00	2.00
17	1.00	1.00	1.00
18	2.00	3.00	4.00

Using Haye's K alpha syntax, the following command line is used:

KALPHA judges = r1 r2 r3 /level = 1/detail = 1/boot = 0.

The three judges are raters 1, 2, and 3, denoted in SPSS as r1, r2, and r3. Level = 1 which means these are nominal scaled ratings (categorical), and detail is 1 means calculations should be reported. Book = 0 means no bootstrapping is to occur.

```
Run MATRIX procedure:

Krippendorff's Alpha Reliability Estimate


               Alpha      Units     Obsrvrs       Pairs
Nominal        .5273    18.0000      3.0000     54.0000

Judges used in these computations:
 r1        r2        r3
```

10. 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.

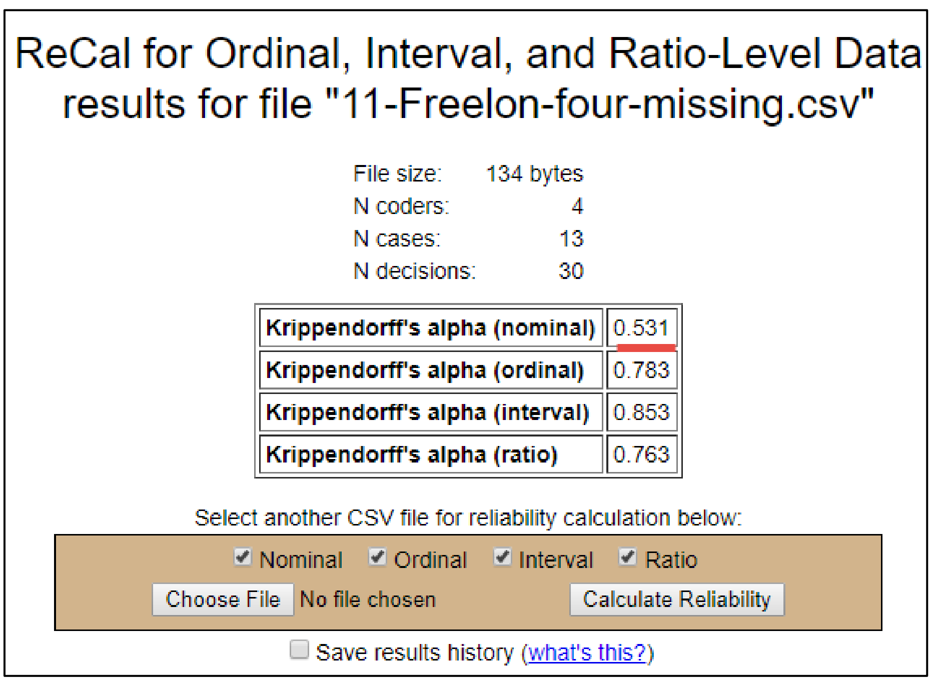
Instructor note: To see difficulties calculating simple percentage agreement with multiple raters and missing data, see three different percent agreement results in this Excel file content/MultipleRatersAgreementPercent.xlsx , three estimates are 72.43%, 65.27%,67.94%, and 63.63%, none of which agree with Geertzen's value of 58.3%)

10a. Freelon's site 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.

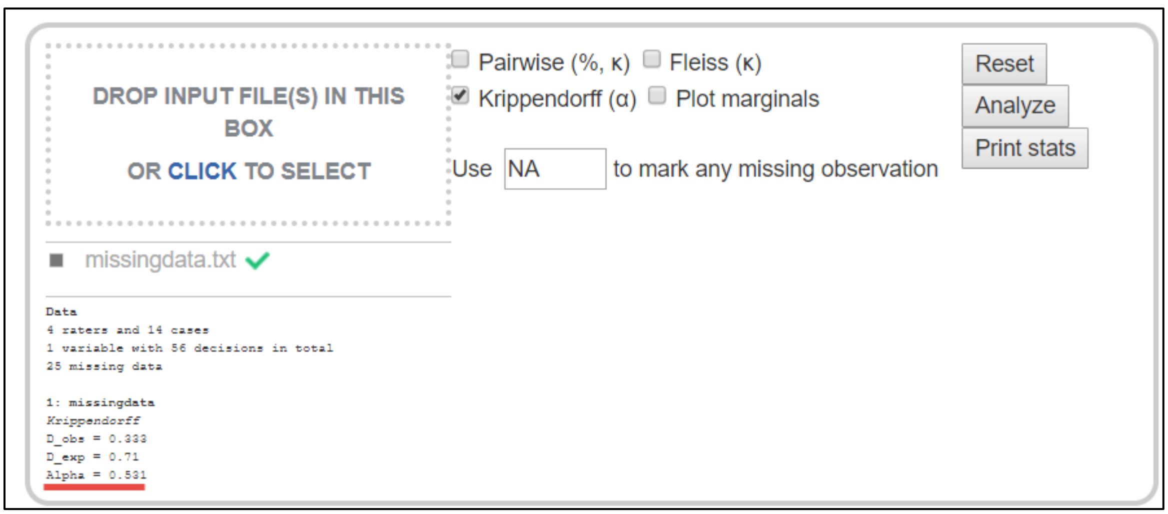


~~10b. Geertzen's site https://nlp-ml.io/jg/software/ira/~~

~~Geertzen's site can be used to find Krippendorff's alpha. To identify missing data, Geertzen requires that missing data be denoted with NA (capital NA, "na" won't work). Below is a revised table to meet Geertzen's specifications.~~

~~Rater 1~~	~~Rater 2~~	~~Rater 3~~	~~Rater 4~~
~~1~~	~~2~~	~~1~~	~~NA~~
~~1~~	~~2~~	~~NA~~	~~NA~~
~~NA~~	~~1~~	~~1~~	~~1~~
~~1~~	~~NA~~	~~NA~~	~~NA~~
~~1~~	~~1~~	~~2~~	~~1~~
~~2~~	~~NA~~	~~2~~	~~NA~~
~~NA~~	~~1~~	~~NA~~	~~1~~
~~2~~	~~NA~~	~~3~~	~~NA~~
~~NA~~	~~2~~	~~2~~	~~NA~~
~~3~~	~~NA~~	~~NA~~	~~3~~
~~3~~	~~NA~~	~~NA~~	~~2~~
~~NA~~	~~NA~~	~~1~~	~~1~~
~~4~~	~~NA~~	~~NA~~	~~4~~
~~4~~	~~4~~	~~NA~~	~~NA~~

~~Results of Geertzen's calculations are presented below. K alpha = .531. The page won't calculate alpha if other statisitics are requested (e.g., Pariwise or Fleiss).~~



10c. SPSS

The SPSS syntax by Hayes also produces the same value of K alpha. See below. Leave missing data as blank in the SPSS data sheet – see example below.

	r1	r2	r3	r4
1	1.00	2.00	1.00	.
2	1.00	2.00	.	.
3	.	1.00	1.00	1.00
4	1.00	.	.	.
5	1.00	1.00	2.00	1.00
6	2.00	.	2.00	.
7	.	1.00	.	1.00
8	2.00	.	3.00	.
9	.	2.00	2.00	.
10	3.00	.	.	3.00
11	3.00	.	.	2.00
12	.	.	1.00	1.00
13	4.00	.	.	4.00
14	4.00	4.00	.	.
15				

Output from Hayes' k-alpha syntax appears below.

```
Run MATRIX procedure:

Krippendorff's Alpha Reliability Estimate

            Alpha      Units     Obsrvrs      Pairs
Nominal     .5307    13.0000      4.0000    22.0000

Judges used in these computations:
 r1        r2        r3        r4
```

Supplemental: For any with access to Stata, here's the command and output to obtain K-alpha with Stata.

First, perform a search for the kalpha command, then download and install. Once installed use this command:

. kalpha var1 var2 var3 var4, scale(n) **transpose** bootstrap(reps(5000) minalpha(.8) dots(10))

Krippendorff's Alpha-Reliability
(nominal data)

No. of units = 13
No. of observers = 4
Krippendorff's alpha = 0.531

Bootstrap results

No. of coincidences = 30
Replications = 5000

[95% Conf. Interval]
0.343 0.718

Probability of failure to reach alpha
min. alpha q
0.800 0.999

Assumes columns are cases and rows coders, so use **transpose** if columns are coders and rows are cases.

11. 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 presented below.

Average Pairwise Percent Agreement

Average pairwise percent agr.	Pairwise pct. agr. cols 1 & 4	Pairwise pct. agr. cols 1 & 3	Pairwise pct. agr. cols 1 & 2	Pairwise pct. agr. cols 2 & 4	Pairwise pct. agr. cols 2 & 3	Pairwise pct. agr. cols 3 & 4
95%	95%	95%	95%	95%	95%	95%

Fleiss' Kappa

Fleiss' Kappa	Observed Agreement	Expected Agreement
0.316	0.95	0.927

Average Pairwise Cohen's Kappa

Average pairwise CK	Pairwise CK cols 1 & 4	Pairwise CK cols 1 & 3	Pairwise CK cols 1 & 2	Pairwise CK cols 2 & 4	Pairwise CK cols 2 & 3	Pairwise CK cols 3 & 4
0.244	-0	0.487	0.487	-0	0.487	-0

Krippendorff's Alpha (nominal)

Krippendorff's Alpha	N Decisions	$\Sigma_c o_{cc}$***	$\Sigma_c n_c(n_c - 1)$***
0.325	80	76	5852

***These figures are drawn from Krippendorff (2007, case C.)

12. Patterns of Response, Bias in Coding Categories, Kappa Paradoxes

This section is under development and not yet ready for use.

Joyce (2013) presents the following tables

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

Figure 5: The Weakness of Cohen's Kappa

Coder A

Categories: Coder B

	a	b	c	
a	12	9	9	30
b	9	14	9	32
c	9	9	20	38
	30	32	38	100

A_0 = .460
π = .186
κ = .186

	a	b	c	
a	12	18	18	48
b	0	14	18	32
c	0	0	20	20
	12	32	56	100

A_0 = .460
π = .186
κ = .258

Percent agreement = 46.0% Scott's pi = .186 Cohen kappa = .186 K alpha for first table = .1836	Percent agreement = 46.0% Scott's pi = .186 Cohen kappa = .258 K alpha for first table = .1898

Note how kappa is influenced by the pattern of response whereas neither pi nor alpha are affected or greatly affected.

Stata output for K alpha (same results for both tables):

. kalpha var1 var2, scale(n) **transpose**

Krippendorff's Alpha-Reliability
(nominal data)

```
      No. of units = 100
   No. of observers = 2
Krippendorff's alpha = 0.190
```

Example tables of paradoxes for kappa: http://folk.ntnu.no/slyderse/Pres24Jan2014.pdf
(in folder as 2014 Lydersen Paradoxes with Agreement Measures.pdf)

12. Instructor notes for content to review

Instructor note: Add the following –
(a) Nominal – Gwet's gamma or AC1 (seems to address some of the difficulties noted with kappa), conditional agreement (Rosenfield et al 1986), Aickin's alpha
(b) Ordinal – weighted kappa for ordered categories, tetrachoric correlation for binary-ordered ratings, polychoric correlation for ordinal ratings (http://www.john-uebersax.com/stat/tetra.htm); if variable has 5 or more ranked categories, consider using Interval or Ratio procedures below.
(c) Interval or Ratio – ICC just like with test-retest reliability (focus on agreement); Cronbach's alpha (focus on reliability), Bland-Altman plot for comparing rating methods/scales rather than raters, factor analysis for interval/ratio (and Likert)-type data (see http://www.john-uebersax.com/stat/cont.htm) also see paragraph on interpretation of factor loadings (interesting perspective on correlation lack of agreement is useful, use multiple indices of agreement and consistency to assess data. “"There is growing awareness that rater agreement should be viewed as having distinct components, and that these components should be assessed distinctly, rather than combined into a single omnibus index. To this end, a statistical modeling approach to such data has been advocated (Agresti, 1992; Uebersax, 1992)."

Instructor note: How to handle missing code for percent agreement and kappa (i.e., one coder provides code, second does not)? Inventing 5th coding option to signal this discrepancy changes kappa but not percent agreement and adds additional category to contingency table which alters calculations.

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

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology. Second Edition*. Thousand Oaks, CA: Sage.

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

*Moore, M. T., & Griffin, B. W. (2006). Identification of factors that influence authorship name placement and decisions to collaborate in peer-reviewed, education-related publications. Studies in Educational Evaluation, 32(2), 125-135.

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=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