

## 07b Coder/Rater Agreement for Ranked Data (Ordinal and Interval/Ratio Data)

### Topics

1. Why Assess Agreement among Coders?
  2. Data: Ordinal, Interval, and Ratio
  3. Agreement vs. Consistency (Reliability)
  4. Ordinal Rating Data
  5. Measures of Agreement and Consistency for Ordinal Data: Focus on Krippendorff's Alpha
  6. Measures of Agreement and Consistency for Ordinal+, Interval, and Ratio Data
  7. Two Raters Example for Ordinal+, Interval, and Ratio Data: Professional Learning Communities
- Stop reading at page 19 – material beyond this page requires review and revision.**

### 1. Why Assess Agreement among Coders?

As explained in the presentation notes for assessing coder agreement for nominal data, 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). Given this, it is important that coding of textual data be done by more than one coder/rater, and that their codes be compared to assess level of agreement.

### 2. Data: Ordinal, Interval, and Ratio

The previous set of notes explained how to assess coder agreement for nominal, or categorical, data. For example, one may be asked to code whether news reports represent one of four categories: political, economic, social, or sports. There is no inherent rank to these categories, so the measures of agreement previously studied would be suitable for assessing agreement among raters assigning these four codes.

For two raters, the following measures of agreement were presented:

- Percent Agreement
- Cohen's kappa
- Scott's pi (very similar to Cohen's kappa)
- Krippendorff's alpha (most flexible measure of agreement available)

When more than two raters are present, the following measures of agreement were considered:

- Mean Percent Agreement
- Mean Cohen's kappa (mean of all pairwise kappa values)
- Fleiss' kappa (extension of Scott's pi for more than two raters)
- Krippendorff's alpha

For the measures of agreement presented below, codes are assumed to represent ranked information. For example, readers may rate essays using the following four categories:

- 1 = fail
- 2 = marginally pass
- 3 = pass
- 4 = superior

These data form an ordinal scale such that categories can be ranked from better to worse.

As another example, a panel of judges may be asked to watch several short videos of police interaction with demonstrators and count the number of times police strike demonstrators. This count from each judge is a ratio variable. Of interest is the degree to which panel members present the same or similar counts. Similar examples could be found for education such as the count of distracting behaviors by students in a classroom for a day, or the number of times a teacher calls on a student or group of students in an hour.

### 3. Agreement vs. Consistency (Reliability)

Recall the comparison between agreement and consistency in the presentation on Test-retest Reliability. The logic for assessment of inter-rater reliability, or agreement, is the same, except instead of focusing upon consistency between two or more time periods, the focus is between two or more raters.

**Consistency** refers to the relative position of scores across two or more raters/coders. Consistency is an assessment of whether coders' scores tend to rank order something in similar positions.

**Agreement** refers to the degree to which two or more raters/coders agree or show little difference in actual scores; the lower the absolute difference, the greater the agreement between coders.

Table 1 below is a reproduction of a table provided in the test-retest notes, except that references to tests are changed to raters/coders. In Table 1 note that for Consistency, the two raters provide scores that show large differences between scores (e.g., for the first data source, rater 1 scores it 95 while rater 2 scores it 44). Despite these large score differences, the rank order of data sources by the two raters is identical, so both raters demonstrate high consistency in rating the data sources.

Table 1 also shows an assessment of agreement in scores between rater 1 and 2 – the difference between the two ratings. The smaller the difference, the higher will be agreement.

Table 1: Relative vs. Absolute Reliability for Ratings from Two Raters

Data Source	Relative Reliability, Consistency				Absolute Reliability, Agreement		
	Rater 1	Rank 1	Rater 2	Rank 2	Rater 1	Rater 2	Difference
1	95	1	44	1	95	92	3
2	90	2	22	2	90	91	-1
3	85	3	20	3	85	83	2
4	80	4	19	4	80	79	1
5	75	5	10	5	75	78	-3
6	70	6	9	6	70	72	-2
7	65	7	8	7	65	64	1
8	60	8	1	8	60	61	-1

As Table 1 demonstrates, it is possible to have consistency without agreement, but it is usually rare to find examples of agreement without consistency. Liao, Hunt, and Chen (2010) argue that one may have high agreement with low consistency and attempt to produce artificial data to support this claim. However, their data (see Table 2, p. 615) does not support their claim of high agreement and low consistency because measures of agreement are all extremely low or negative (K alpha = -.29 and ICC for agreement = -.38 and -.500). Additionally, Tinsley and Weiss (1975) make this claim, but their data (see their Table 1) also fails to demonstrate agreement (k alpha = -.12 ordinal or -.098 interval).

Which do we use for assessing reliability among raters, consistency or agreement? In most cases researchers are interested in agreement – showing that raters provide the same or similar scores. The pattern of ratings is usually not relevant, so rater consistency is of little interest. However, in some situations one desires a measure of consistency. For example, if raters are asked to independently develop scales and rate something, such as observed anti-social behavior,

one would be interested not in agreement since it is unlikely the two raters develop identical rating scales, instead, one would be interested in knowing whether the two raters independently produce consistent ratings (i.e., similar high and low assessments of sampled anti-social behavior).

#### 4. Ordinal Rating Data

We understand ordinal to mean a variable that has mutually exclusive categories with a natural rank to those categories. Below are rating scales that represent ordinal data.

##### (a) Binary Classification

Below are examples of binary scales used to rate data.

High	Good	Pass	Hard	Hot
Low	Bad	Fail	Soft	Cold

While such rating scales can be classified as ordinal, there is nothing to be gained, statistically, from the ordinal ranking when only two classification options are present, so the agreement measures presented earlier for nominal data could be used for these types of ratings (e.g., percent agreement, Cohen's kappa, Scott's pi, Fleiss's kappa, Krippendorff's alpha).

##### (b) Three Categories

Below are examples of ordinal classification scales with three options.

Excellent	Above Average	Hot
Acceptable	Average	Warm
Unacceptable	Below Average	Cold

With a three-step rating scale order can make a difference in assessment of agreement among raters when compared to agreement measures used for nominal data. For example, two judges are asked to read essays and assign one of three scores as outlined below.

- 3 = Excellent
- 2 = Acceptable
- 1 = Unacceptable

With a measure of agreement that assumes nominal data, the difference between a rating of 3 and a rating of 1,  $3 - 1 = 2$  is the same as the difference between a rating of 3 and a rating of 2,  $3 - 2 = 1$  because with nominal data ranking and order is meaningless. When codes are of the nominal scale, then numbers 1, 2, and 3 are simply labels like the labels orange, apple, and grape, or green, blue, and red, there is no natural rank.

However, with ordinal data, the numeric differences among ratings take meaning. The closer the numbers, the closer the raters, hence a difference of  $3 - 2 = 1$  means the two raters are closer in agreement than a difference of  $3 - 1 = 2$ . Given this, measures of agreement for ordinal, and interval and ratio, data should take into account how close raters are in agreement.

As an illustration, below are fictional essay ratings from two judges. The three-step scale presented above (3 = Excellent, etc.) is used.

Student Essay	Judge 1	Judge 2
1	3	3
2	1	2
3	2	3
4	2	2
5	1	1
6	3	2
7	3	2

Krippendorff's alpha can be used for data of any scale (nominal, ordinal, interval, or ratio). Below are results showing Krippendorff's alpha for the four scales of measurement for these data. Note that as the scale of measurement is refined (going from nominal to ordinal, or ordinal to interval, or interval to ratio), alpha grows in strength for these three categories. This shows that Krippendorff's alpha is sensitive to size of the difference among rating scores, and this sensitivity grows as the scale of measurement is refined.

<b>Krippendorff's alpha (nominal)</b>	0.175
<b>Krippendorff's alpha (ordinal)</b>	0.479
<b>Krippendorff's alpha (interval)</b>	0.519
<b>Krippendorff's alpha (ratio)</b>	0.568

### (c) Four Categories

Three-step scales can be converted to four-step scales easily. For example:

Excellent	Superior	Hot
Acceptable	Above Average	Warm
Marginally Acceptable	Below Average	Lukewarm
Unacceptable	Poor	Cold

The procedures discussed below for assessing agreement for three ranked categories will also work for assessing four ranked categories.

### (d) Five or More Categories

When scales have five or more steps, I recommend treating them as interval data and use procedures discussed below for interval and ratio data. This assumes the scale rating steps appear to form an approximately equally spaced continuum (e.g., similar to Likert-type scales that range from Strongly Disagree to Strongly Agree with several categories between these ends).

This recommendation is not universally held since some argue that ordinal data are not interval and therefore should not be treated as interval. I have found that ordinal data with five or more categories tend to work well in analysis procedures that assume interval or ratio data, and this is especially true when multiple ordinal items are combined to form composites (e.g., measurement scales that employ multiple indicators that are combined).

## 5. Measures of Agreement and Consistency for Ordinal Data: Focus on Krippendorff's Alpha

There are several measures of agreement and consistency for ordinal data, and these include:

### Agreement

- Krippendorff's alpha
- Cohen weighted kappa (not covered here)
- Brennan-Prediger kappa (not covered here)
- Fleiss's kappa with weights (not covered here)

### Consistency

- tetrachoric correlation for binary-ordered ratings
- polychoric correlation for ordinal ratings

To be reviewed and possibly added in the future:

- Gwet's (2014) AC1 (or gamma,  $\Upsilon$ )
- Gwet's (2014) AC2 (or gamma,  $\Upsilon$ )

**(Instructor's Note** – add discussion for assessing consistency of ordinal data; when and how to assess)

Unfortunately, none, or few, of the above measures are implemented in SPSS. Given this, we will rely on Freelon's site to calculate Krippendorff's alpha for ordinal data.

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

Krippendorff's alpha is one of the better measures of **agreement**, works well for missing data, can be used for 2 or more raters, and works for all four scales of measurement.

### Example 1: Judges Essay Scores

For this example, use the data presented above for judges rating essays according to a three-point scale, shown below.

- 3 = Excellent
- 2 = Acceptable
- 1 = Unacceptable

Steps for finding Krippendorff's alpha are illustrated below.

#### Example 1-1. Prepare data for uploading to Freelon's site

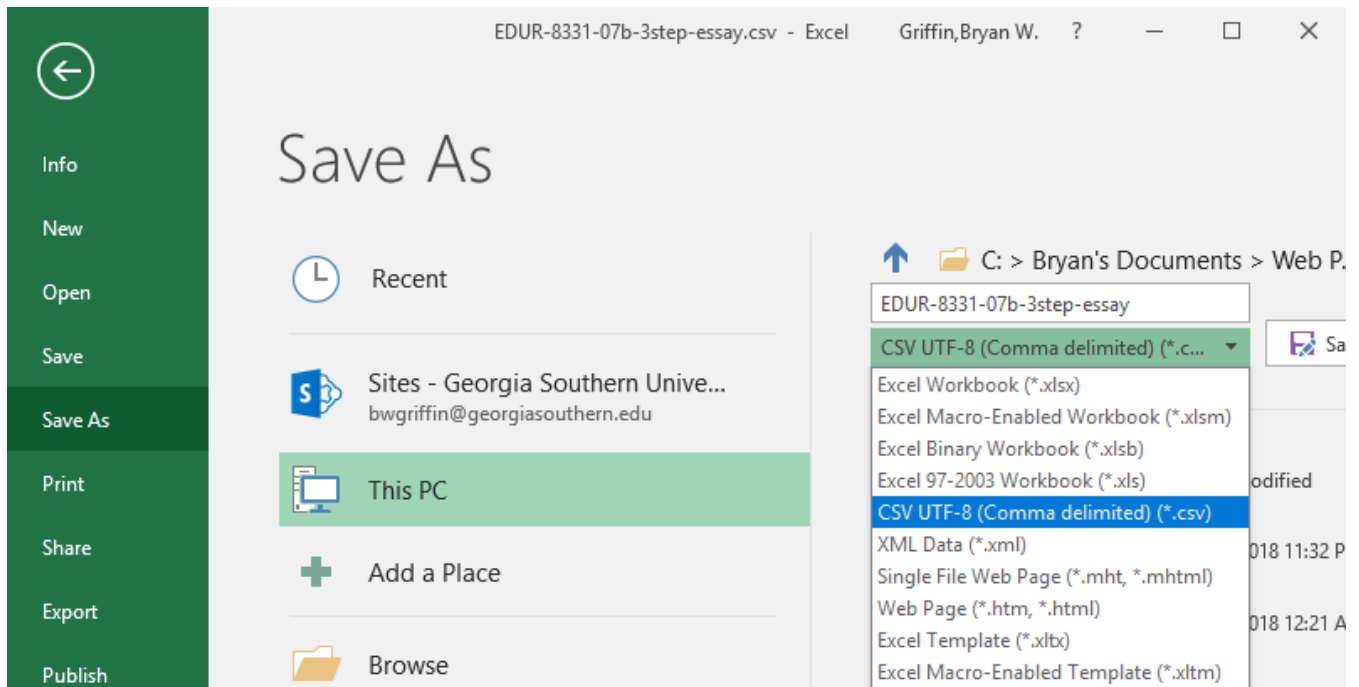
Enter these data into Excel, Google Sheet, or some CSV (comma separated values, or comma delimited file) producing software.

Student Essay	Judge 1	Judge 2
1	3	3
2	1	2
3	2	3
4	2	2
5	1	1
6	3	2
7	3	2

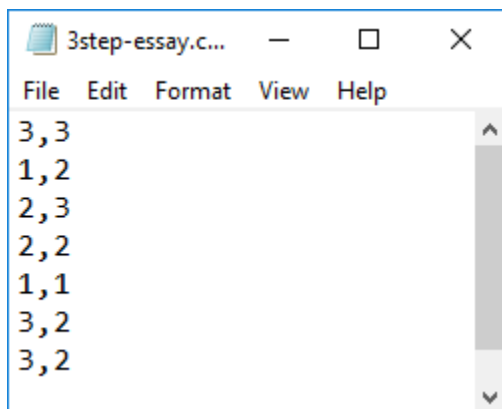
Excel entry is shown below. Note that there is no header and only the ratings are presented. Include no other information, like the essay identification column shown above, otherwise calculations will fail or be incorrect.

	A	B	
1	3	3	
2	1	2	
3	2	3	
4	2	2	
5	1	1	
6	3	2	
7	3	2	
8			

### Example 1-2. Save Data in Comma Separated Values format, CSV



When viewed in Notepad, the file should look like the image below. Note there are no other data entered – only the ratings for each rater separated by commas.



### Example 1-3. Upload data to Freelon's site

Open Freelon's site, then select his ReCal OIR page.

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

dfreelon.org

search this site Q

home
cv/pubs
research resources
utilities
blog
contact
calculate intercoder reliability

## ReCal: reliability calculation for the masses

UPDATE 5/22/17: By popular demand, [ReCal OIR](#) now allows missing data! [Click the link for details.](#)

ReCal ("Reliability Calculator") is an online utility that computes intercoder/interrater reliability coefficients for [nominal](#), [ordinal](#), [interval](#), or [ratio-level](#) data. It is compatible with Excel, SPSS, STATA, OpenOffice, Google Docs, and any other database, spreadsheet, or statistical application that can export comma-separated ([CSV](#)), tab-separated ([TSV](#)), or semicolon-delimited data files.

ReCal consists of three independent modules each specialized for different types of data. The following table will help you select the module that best fits your data. (If you do not know whether your data are considered nominal, ordinal, interval, or ratio, please consult [this Wikipedia article](#) to find out more about these levels of measurement.)

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

Please visit the [ReCal FAQ/troubleshooting page](#) if you have questions or are experiencing difficulty getting ReCal to work with your data. If you still have questions please [contact me directly](#), rather than leaving a comment.

Want to support ReCal? The best way is with a citation to one or both of the following

#### Categories

campaign 2012 charts communication computational social science conference musings fyi gephi internet & politics net politics online deliberation race **recal scholarly tools** social network analysis twitter uncategorized updates

#### Recent Posts

- ▶ Beyond the Hashtags Twitter data
- ▶ Social media collection tools: A curated list
- ▶ Comm depts.: Want to excel in computational methods? Do these four things.
- ▶ Co-citation map of 9 comm journals, 2003-2013
- ▶ T2G 0.3: Visualize only RTs or mentions in Gephi

The next page that opens for the ReCal OIR link above is shown below.

**dfreelon.org**  
Deen Freelon, associate professor, UNC School of Media and Journalism

home cv/pubs research resources utilities blog contact [calculate intercoder reliability](#)

## ReCal for Ordinal, Interval, and Ratio Data (OIR)

**UPDATE 5/22/17:** By popular demand, ReCal OIR now allows missing data! [See documentation below for details.](#)

ReCal OIR ("Reliability Calculator for Ordinal, Interval, and Ratio data") is an online utility that computes intercoder/interrater reliability coefficients for [nominal, ordinal, interval, and ratio](#) data judged by **two or more coders**. (If you need to calculate reliability for nominal data judged by two coders only, use [ReCal2](#); for nominal data coded by three or more coders, use [ReCal3](#). *As of 5/22/17, ReCal OIR can also be used to compute coefficients for incomplete nominal datasets.*) Here is a brief feature list:

- Calculates **three** ~~four~~ reliability coefficients:
  - Krippendorff's alpha for nominal data
  - Krippendorff's alpha for ordinal data
  - Krippendorff's alpha for interval data
  - Krippendorff's alpha for ratio data
- Accepts any range of possible variable values, including decimal values and negative numbers
- Allows missing data (*as of 5/22/17*)
- Results should be valid for **nominal, ordinal, interval, or ratio data sets coded by two or more coders** (other uses are not endorsed, and accurate results are not guaranteed in any case — trust but verify!)

If you have used ReCal OIR before, you may submit your data file for calculation via the form below. If you are a first-time user, please read [the documentation](#) first. (*Note: failure to format data files properly may produce incorrect results!*) You should also read ReCal's [very short license agreement](#) before use.

Nominal
  Ordinal
  Interval
  Ratio

No file chosen

### Categories

campaign 2012 charts communication computational social science conference musings fyi gephi internet & politics net politics online deliberation race

**recal scholarly tools** social network analysis twitter uncategorized updates

### Recent Posts

- Beyond the Hashtags Twitter data
- Social media collection tools: A curated list
- Comm depts.: Want to excel in computational methods? Do these four things.
- Co-citation map of 9 comm journals, 2003-2013
- T2G 0.3: Visualize only RTs or mentions in Gephi

Click on "Choose File" and upload the CSV file created for the essay data.

The screenshot shows a file selection dialog box titled "Open" with the following content:

- Path: Web Pages > Courses > edur8331 > edur8331-presentations
- Files listed:
 

Name	Date modified
EDUR-8331-07b-coder-agreement-ranked-codes.docx	10/19/2018 12:21
EDUR-8331-07b-3step-essay.csv	10/19/2018 12:21
~SUR-8331-07a-coder-agreement-nominal-data.docx	10/18/2018 11:43
~SUR-8331-07b-coder-agreement-ranked-codes.docx	10/18/2018 11:40
- File name field: Empty
- File type: All Files
- Buttons: Open, Cancel

A red arrow points from the "Choose File" button in the background interface to the "EDUR-8331-07b-3step-essay.csv" file in the dialog.

Next place a mark next to "Ordinal" then click on "Calculate Reliability" to obtain results.



**ReCal for Ordinal, Interval, and Ratio-Level Data**  
results for file "3step-essay.csv"

File size: 38 bytes  
N coders: 2  
N cases: 7  
N decisions: 14

**Krippendorff's alpha (ordinal)** 0.479

Select another CSV file for reliability calculation below:

Nominal  Ordinal  Interval  Ratio

No file chosen

Save results history ([what's this?](#))

Krippendorff's alpha for these data is .479.

### Interpretation of Krippendorff's alpha

Krippendorff (2004) wrote the following:

<b>.80 or higher</b>	"An acceptable level of agreement below which data are to be rejected as too unreliable..." [in serious situations such as legal issues, human lives at stake, etc.]
<b>.667 or higher</b>	"where tentative conclusions are still acceptable"

Given these guidelines, the alpha of .479 suggests the raters have too little agreement to be considered reliable.

### Example 2: Four Raters

As another example, fictitious data will be used to assess the level of agreement among four evaluators who are asked to rate grant applications. The scale is four steps, as shown below.

- 4 = Superior, certainly fund grant
- 3 = Above Average, fund grant if money available
- 2 = Below Average, do not fund grant
- 1 = Poor, do not fund grant

The data appear below. Note that one score is missing from Rater 3 for application 7.

Grant Applications	Rater 1	Rater 2	Rater 3	Rater 4
1	4	3	2	3
2	3	3	3	3
3	2	1	2	2
4	4	2	3	3
5	1	2	2	1
6	1	1	1	1
7	2	1		1
8	3	3	4	4

For missing data, Freelon's site requires that the hashtag symbol be inserted in the missing data space, like shown below in the Excel file. This is critical otherwise that row of data will not be included, or an error message will appear.

	A	B	C	D	E
1	4	3	2	3	
2	3	3	3	3	
3	2	1	2	2	
4	4	2	3	3	
5	1	2	2	1	
6	1	1	1	1	
7	2	1	#	1	
8	3	3	4	4	
9					
10					

Here are the comma delimited data shown in Notepad with the hashtag showing.

```

grant-ratings.csv
File Edit Format View Help
4,3,2,3
3,3,3,3
2,1,2,2
4,2,3,3
1,2,2,1
1,1,1,1
2,1,#,1
3,3,4,4

```

Results of the analysis re shown below.

### ReCal for Ordinal, Interval, and Ratio-Level Data results for file "grant-ratings.csv"

File size: 75 bytes  
N coders: 4  
N cases: 8  
N decisions: 31

**Krippendorff's alpha (ordinal)** 0.725

Select another CSV file for reliability calculation below:

Nominal    Ordinal    Interval    Ratio

No file chosen

Alpha = .725 which is acceptable in this situation.

## 6. Measures of Agreement and Consistency for Ordinal+, Interval, and Ratio Data

As noted above, if the rating scale is 5 or more steps and appears to form an approximately equally spaced continuum, or if the data are clearly interval or ratio, then one may choose from a number measures of consistency and agreement.

Above I explained the distinction between consistency and agreement, and argued that for assessing raters, agreement is often the measure that should be sought. Uebersax argues that consistency and agreement are two dimensions of the data, and both should be assessed.

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

**(Instructor's note** – expand this discussion)

There are several measures of agreement and consistency for ordinal+, interval, and ratio data, and these include:

### Agreement

- Intraclass Correlation Coefficient (ICC), agreement model
- Krippendorff's alpha for ordinal, interval, and ratio data

### Consistency

- Pearson Correlation
- Cronbach's alpha
- Intraclass Correlation Coefficient (ICC), consistency model
- Factor Analysis

To be reviewed and possibly added in the future:

- Gwet's (2014) AC1 (or gamma,  $\Upsilon$ )
- Gwet's (2014) AC2 (or gamma,  $\Upsilon$ )

## 7. Two Raters Example for Ordinal+, Interval, and Ratio Data: Professional Learning Communities

Suppose two raters are asked to rate 10 high schools in terms of level of integration for Professional Learning Communities (PLC). The scale ranges from low of 1 to high of 10. Below are their ratings.

High School	Rater 1	Rater 2
1	1	5
2	3	7
3	4	4
4	6	7
5	2	5
6	8	9
7	10	10
8	7	8
9	4	7
10	5	8

## 7.1 Measures of Agreement

### 7.1a Krippendorff's Alpha

Krippendorff's alpha can be calculated for these data. Using the steps outlined above, a CSV file was created for the PLC data and uploaded to Freelon's site. Results are presented below. Perform this analysis yourself to ensure you can replicate the results shown below.

### ReCal for Ordinal, Interval, and Ratio-Level Data results for file "PLC-two-raters.csv"

File size: 75 bytes  
N coders: 2  
N cases: 10  
N decisions: 20

<b>Krippendorff's alpha (ordinal)</b>	0.538
<b>Krippendorff's alpha (interval)</b>	0.517
<b>Krippendorff's alpha (ratio)</b>	0.221

Select another CSV file for reliability calculation below:

Nominal  
 Ordinal  
 Interval  
 Ratio

No file chosen  

Save results history ([what's this?](#))

The question here is which measure should be use? Since Krippendorff allows one to make a distinction among ordinal, interval, and ratio data, and since these data were derived, most likely, from an ordinal classification system, that is measure that should be most accurate, and interval would be next. Since the data are less precise than that required for ratio scale, it should not be used.

According to these results, agreement is not strong with estimates of .538 and .517.

### 7.1b Intraclass Correlation Coefficient, ICC

Recall discussion of the ICC for using with test-retest reliability. To use ICC for rater reliability, one must first determine how raters were selected, whether one wishes to assess consistency or agreement, and whether one wishes to know the likely reliability for a single rater or the reliability for several raters when the rating they provide are averaged.

### How were raters/judges selected, Which Case?

Shrout and Fleiss (1979) explain how to select which ICC to use:

Case	Example	ANOVA Model
Case 1 = different judges rate different items; not using the same judges every	Different locations involved in study has different judges, ratings from different locations are pooled so targets rated may not have the same judges. In short, if you don't use the same judges every time, this is a Case 1 situation.	One-way Random
Case 2 = use the same judges, but they were randomly selected from a larger group of judges	There is a pool of 15 individuals who can serve as judges, 5 are randomly selected and those 5 are used to rate all targets.	Two-way Random Effects (both judges and targets are random effects)
Case 3 = use the same judges and they were not randomly selected from a larger pool of judges	Three judges volunteered or were recruited to rate all targets in the study.	Two-way Mixed (judges are a fixed effect and random)

**(Instructor's note** – add discussion of ICC and use of one judge who provides multiple ratings per target, i.e., intra-judge reliability assessment)

In most cases in educational research, the same judges, or nearly the same judges, will be used to provide ratings, so this represents a Case 3 study which uses a two-way mixed ANOVA.

### Agreement or consistency?

If raters used the same rating scale when judging their targets, most likely a measure of agreement is sought. If raters did not use the same scales, then consistency will be sought.

### One Rating or Mean Ratings?

If the goal is to learn how well one rater can evaluate data and produce a reliable score, then **single judge or single measure** should be used (even when the data were evaluated from multiple raters). If the goal is use a mean score from multiple raters, use the **average measures** should be used.

### ICC with SPSS

Below is an illustrated example of ICC with SPSS.

#### (a) Enter Data in SPSS

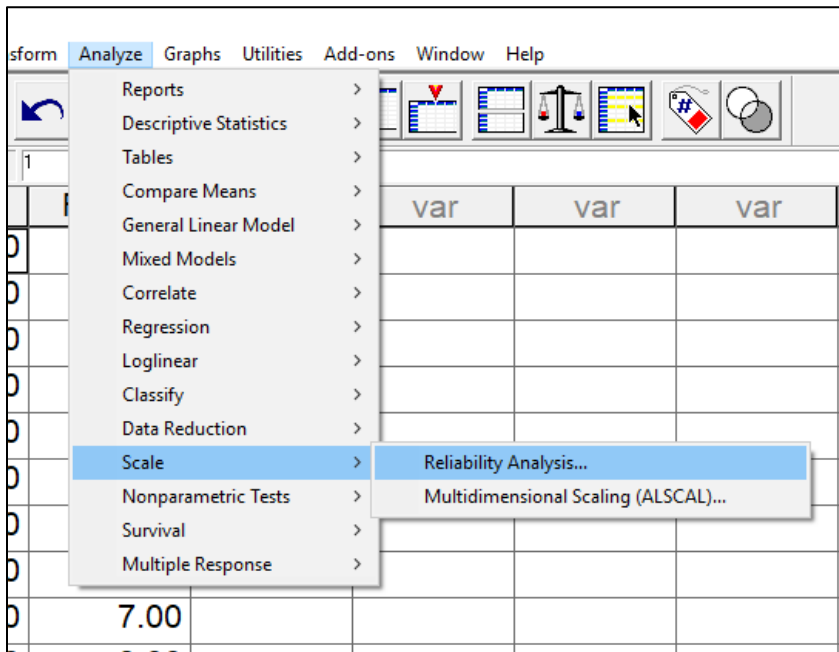
The image below shows the PLC data in SPSS.

	Rater1	Rater2	va
1	1.00	5.00	
2	3.00	7.00	
3	4.00	4.00	
4	6.00	7.00	
5	2.00	5.00	
6	8.00	9.00	
7	10.00	10.00	
8	7.00	8.00	
9	4.00	7.00	
10	5.00	8.00	

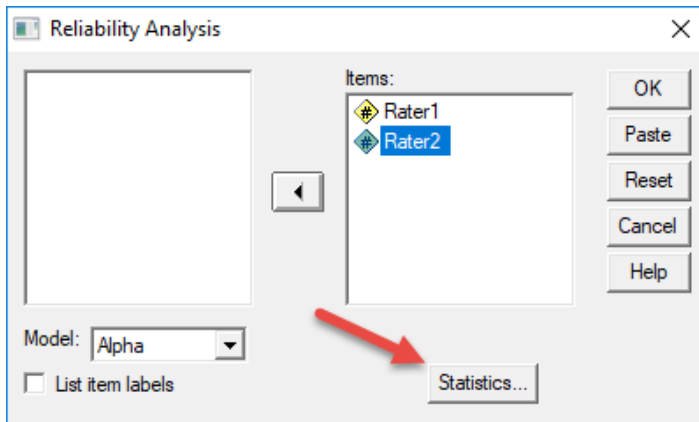
### (b) SPSS Commands

Commands are

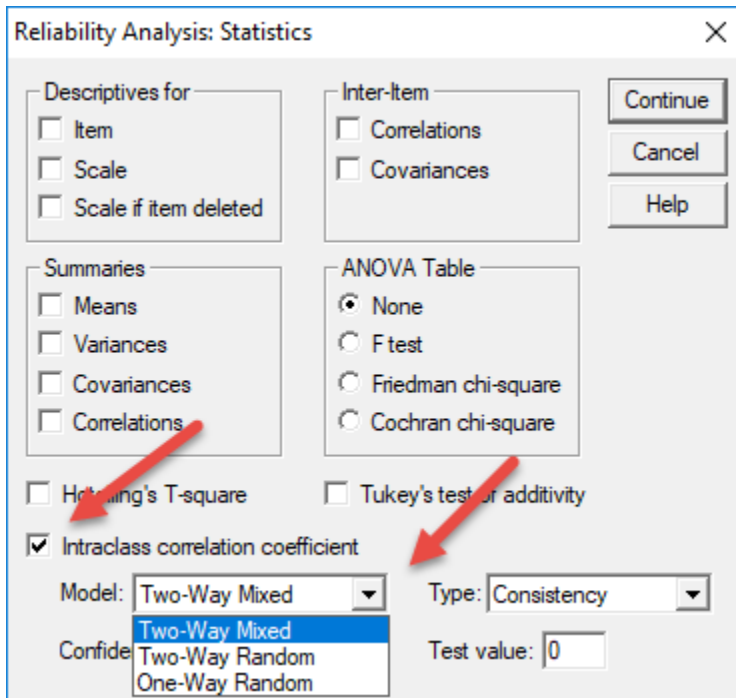
**Analyze -> Scale -> Reliability Analysis**



Move raters' scores from the variable box on the left to the **Items** box on the right, see below.



Select **Statistics** to access the ICC menu, see arrow above. Next, place a mark next to **Intraclass correlation coefficient**, then select the **down arrow button** next to **Model** to access the three model types.



In this example we used the same raters for each school, and we will assume the raters were recruited and are therefore not a random selection from a pool of raters. This is likely the case for most research and evaluation situations. Given this, we will use the **Two-way Mixed** model.

Next, select the **down arrow button** next to Type to access the option between **Consistency** and **Absolute Agreement**; here we want to know level of Agreement. See below.

Reliability Analysis: Statistics

Descriptives for

Item

Scale

Scale if item deleted

Inter-Item

Correlations

Covariances

Summaries

Means

Variances

Covariances

Correlations

ANOVA Table

None

F test

Friedman chi-square

Cochran chi-square

Hotelling's T-square

Tukey's test of additivity

Intraclass correlation coefficient

Model: Two-Way Mixed

Type: Consistency

Confidence interval: 95 %

Test v


Consistency

Absolute Agreement

Continue

Cancel

Help



To obtain results, click on **Continue** and **Ok**.

**Reliability Statistics**

Cronbach's Alpha	N of Items
.879	2

**Intraclass Correlation Coefficient**

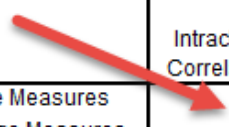
	Intraclass Correlation <sup>a</sup>	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.589 <sup>b</sup>	-.099	.891	8.273	9.0	9	.002
Average Measures	.741 <sup>c</sup>	-.291	.944	8.273	9.0	9	.002

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.



Results are shown above. There are two ICC values reported, one for **Single Measures**, .589, and one for **Average Measures**, .741. In this case the single measure tells us if we used just one rater, the reliability would be low, which is similar to the Krippendorff alpha of .538.

If we planned to take the average of ratings from multiple judges, the reliability would be .741, which is higher, but still shows too little agreement among judges.



## Interpretation of ICC

Koo and Li (2016) suggest not relying upon the single ICC value, but instead offer the following guide for interpreting the ICC 95% confidence interval (CI).

Use of 95% CI	
<b>.90 or higher</b>	excellent
<b>.75 to .90</b>	good
<b>.50 to .75</b>	moderate
<b>.50 or less</b>	poor

Consider the range of possible values provided by the CI since it is more telling than the single point estimate of the ICC.

For example, SPSS reports the **95% Confidence Interval** for **single measure** the 95% CI is -.099 to .891. This tells that agreement could be as low as -.09 which is clearly a poor fit, or as high as .891, a good fit.

A similar result was obtained for the **average measure**, with the CI ranging from -.291 to .944. The low of -0.291 suggests no agreement, the high value of .944 suggests excellent fit.

If one of the interval values is negative, that is a clear sign that agreement is lacking. The confidence intervals should not contain 0.00, and typically should be much tighter around the ICC estimate.

**(Instructor's note** – review other recommended interpretations for ICC.)

### 7.1c Agreement Among Three Raters

Extend the Professional Learning Communities (PLC) scenario. Assume there are now three raters as presented below.

High School	Rater 1	Rater 2	Rater 3
1	1	5	1
2	3	7	4
3	4	4	4
4	6	7	9
5	2	5	4
6	8	9	8
7	10	10	6
8	7	8	4
9	4	7	6
10	5	8	6

What are the values for Krippendorff's alpha and ICC? Enter these data to determine whether you can replicate the results I show below.

<b>Krippendorff's alpha (ordinal)</b>	0.495
<b>Krippendorff's alpha (interval)</b>	0.516
<b>Krippendorff's alpha (ratio)</b>	0.437

Intraclass Correlation Coefficient							
	Intraclass Correlation <sup>a</sup>	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.558 <sup>b</sup>	.171	.847	6.870	9.0	18	.000
Average Measures	.791 <sup>c</sup>	.350	.944	6.870	9.0	18	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

### 7.1d Agreement Among Three Raters with Missing Data

High School	Rater 1	Rater 2	Rater 3
1	1	5	1
2	3	7	4
3	4	4	
4	6	7	9
5	2	5	4
6	8	9	8
7	10		6
8	7	8	4
9	4	7	6
10	5	8	6

Find alpha and ICC for these cases with missing data.

Krippendorff's alpha (ordinal)	0.434
Krippendorff's alpha (interval)	0.45
Krippendorff's alpha (ratio)	0.403

ICC with SPSS.

Note that with ICC, the two schools with missing data are omitted from the analysis. This brings our sample from 10 to 8 schools.

Case Processing Summary			
		N	%
Cases	Valid	8	80.0
	Excluded <sup>a</sup>	2	20.0
	Total	10	100.0

a. Listwise deletion based on all variables in the procedure.

Intraclass Correlation Coefficient

	Intraclass Correlation <sup>a</sup>	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.530 <sup>b</sup>	.091	.862	7.667	7.0	14	.001
Average Measures	.772 <sup>c</sup>	.178	.950	7.667	7.0	14	.001

Two-way mixed effects model where people effects are random and measures effects are fixed.

- Type A intraclass correlation coefficients using an absolute agreement definition.
- The estimator is the same, whether the interaction effect is present or not.
- This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Stop reading here – the material requires review and revision.

## 7.2 Measures of Consistency

- Pearson r
- ICC for consistency
- Cronbach's alpha

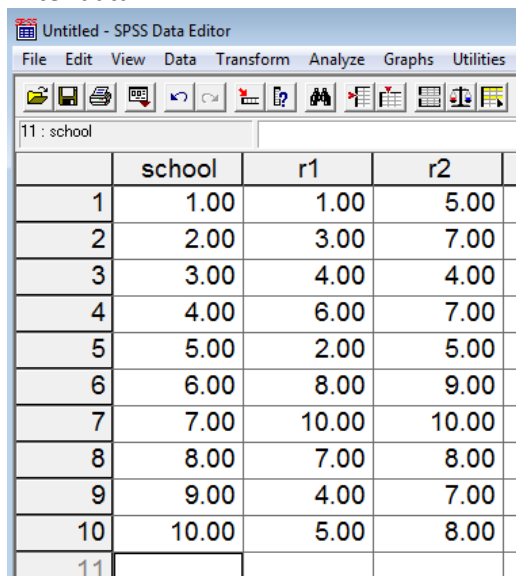
### 7.2.1 Pearson Correlation

Below is an illustration demonstrating why correlation does not assess agreement.

If two raters provide ranked ratings, such as on a scale that ranges from strongly disagree to strongly agree or very poor to very good, sometimes researchers use Pearson's correlation to assess level of agreement between the raters. Pearson's correlation does not measure agreement so it should not be used to assess rater agreement. See illustration below.

### Correlation in SPSS

Enter data:

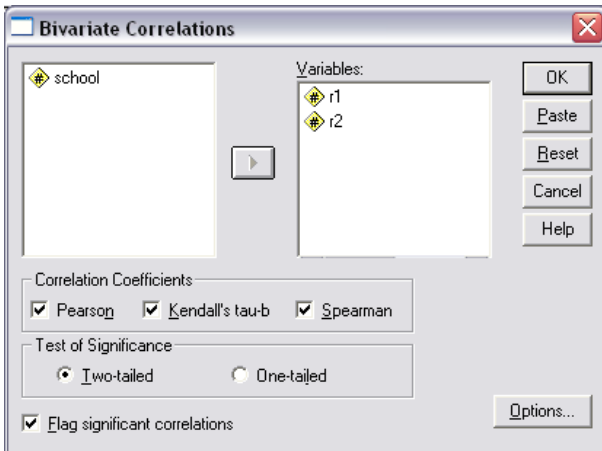


	school	r1	r2
1	1.00	1.00	5.00
2	2.00	3.00	7.00
3	3.00	4.00	4.00
4	4.00	6.00	7.00
5	5.00	2.00	5.00
6	6.00	8.00	9.00
7	7.00	10.00	10.00
8	8.00	7.00	8.00
9	9.00	4.00	7.00
10	10.00	5.00	8.00
11			

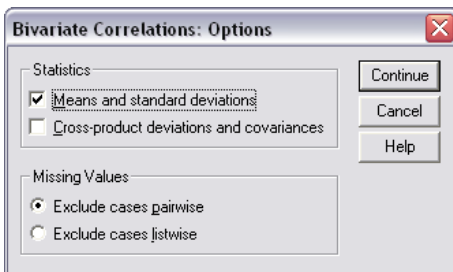
In SPSS, click on

**Analyze → Correlate → Bivariate**

This opens a pop-up window for correlation. Select the two raters and move both to the variable box. Place marks next to Pearson, Kendall's tau-b, and Spearman. See below for an example.



Then selection Options and choose Means and Standard Deviations, then select Continue.



Select "OK" to run the correlation.

**Results**

**Descriptive Statistics**

	Mean	Std. Deviation	N
r1	5.0000	2.78887	10
r2	7.0000	1.88562	10

**Correlations**

		r1	r2
r1	Pearson Correlation	1	.845**
	Sig. (2-tailed)		.002
	N	10	10
r2	Pearson Correlation	.845**	1
	Sig. (2-tailed)	.002	
	N	10	10

\*\* . Correlation is significant at the 0.01 level

The Pearson correlation is  $r = .845$  which suggests high consistency between the two raters, but note that the two means differ, 5.00 vs 7.00 and this indicates that one rater may be rating schools consistently higher than the other rater.

Correlations			r1	r2
Kendall's tau_b	r1	Correlation	1.000	.739(**)
		Coefficient		
		Sig. (2-tailed)	.	.004
	r2	N	10	10
		Correlation	.739(**)	1.000
		Coefficient		
Spearman's rho	r1	Sig. (2-tailed)	.004	.
		N	10	10
		Correlation	1.000	.842(**)
	r2	Coefficient	.842(**)	1.000
		Sig. (2-tailed)	.002	.
		N	10	10

\*\* Correlation is significant at the 0.01 level (2-tailed).

Some researchers and statisticians argue that Pearson's correlation coefficient is inappropriate when data are just ordinal (ranked data) and therefore should not be used. Alternative correlations for ordinal data include Kendall's tau and Spearman's rho.

#### x. Intra-class Correlation Coefficient, ICC

A measure of rater agreement and also rater consistency. One may choose either, but for most cases with raters, agreement is preferred. Unlike Krippendorff's alpha, ICC does not work with missing data – missing cases are deleted listwise.

David Nichols of SPSS explains in the page linked below the difference between ICC with consistency and agreement, and also the various models possible for ICC.

<http://www.ats.ucla.edu/stat/spss/library/whichicc.htm>

ICC ANOVA Models

#### (a) One-way ANOVA Random

Use this approach if one does not know which raters provided which ratings. Normally one will know which raters provided which ratings. For example suppose six raters used, three for each school, but the identities of the raters were unknown, only the 3 ratings per school were provided:

[http://www.bwgriffin.com/gsu/courses/edur9131/temp/inter\\_rater\\_agreement\\_ordinal.sav](http://www.bwgriffin.com/gsu/courses/edur9131/temp/inter_rater_agreement_ordinal.sav)

High School	Rating	Rating	Rating
1	1	5	1
2	3	7	4

3	4	4	4
4	6	7	9
5	2	5	4
6	8	9	8
7	10	10	6
8	7	8	4
9	4	7	6
10	5	8	6

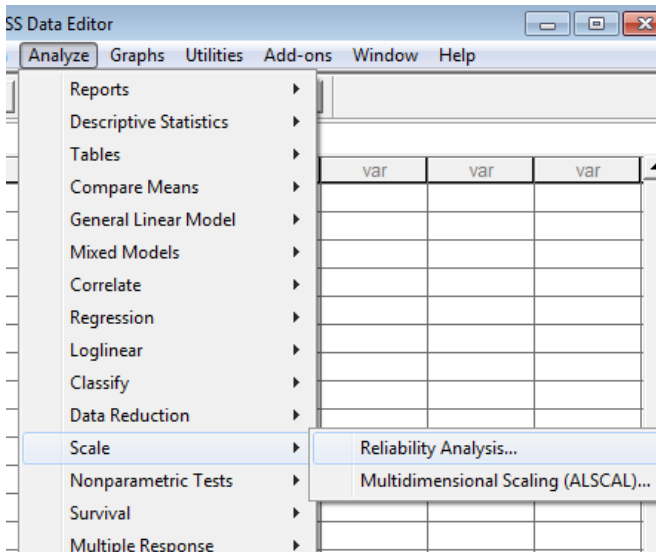
inter\_rater\_agreement\_ordinal - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities

11 : school

	school	rater1	rater2	rater3	va
1	1.00	1.00	5.00	1.00	
2	2.00	3.00	7.00	4.00	
3	3.00	4.00	4.00	4.00	
4	4.00	6.00	7.00	9.00	
5	5.00	2.00	5.00	4.00	
6	6.00	8.00	9.00	8.00	
7	7.00	10.00	10.00	6.00	
8	8.00	7.00	8.00	4.00	
9	9.00	4.00	7.00	6.00	
10	10.00	5.00	8.00	6.00	

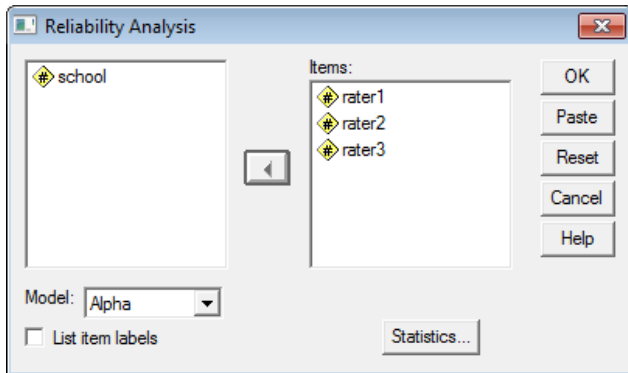
**Analyze → Scale → Reliability Analysis**



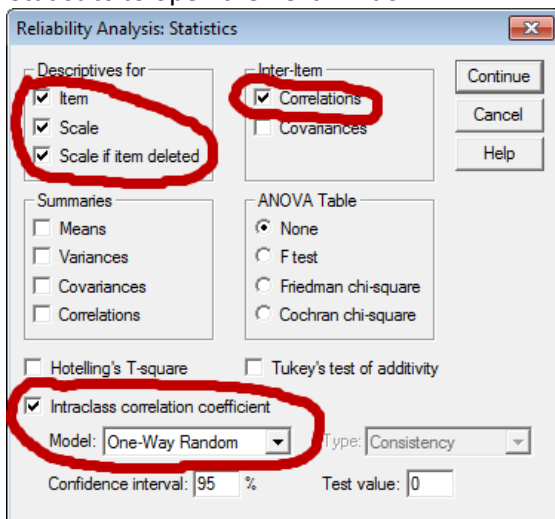
Move the raters from the variables box (not labeled) to the “Items” box. Click on “Statistics” and select the following:

- Correlations
- Scale
- Scale if item deleted
- Intra-class correlation coefficient

See image below as illustration



Click on Statistics to open the next window.



Click "Continue" then "OK" to obtain results. Results are reported below.



Item Statistics

	Mean	Std. Deviation	N
rater1	5.0000	2.78887	10
rater2	7.0000	1.88562	10
rater3	5.2000	2.29976	10

Inter-Item Correlation Matrix

	rater1	rater2	rater3
rater1	1.000	.845	.641
rater2	.845	1.000	.564
rater3	.641	.564	1.000

The covariance matrix is calculated and used in the analysis.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
rater1	12.2000	13.733	.828	.754	.712
rater2	10.2000	21.289	.792	.715	.772
rater3	12.0000	20.222	.634	.413	.879

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
17.2000	38.622	6.21468	3



Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.534	.158	.836	4.439	9	20	.003
Average Measures	.775	.361	.939	4.439	9	20	.003

One-way random effects model where people effects are random.

Single Measures = .534 --- This tells us the expected consistency if one judges provides scores. Note that a value of .534 is not good. Expect that single judges to be less consistent than multiple judges.

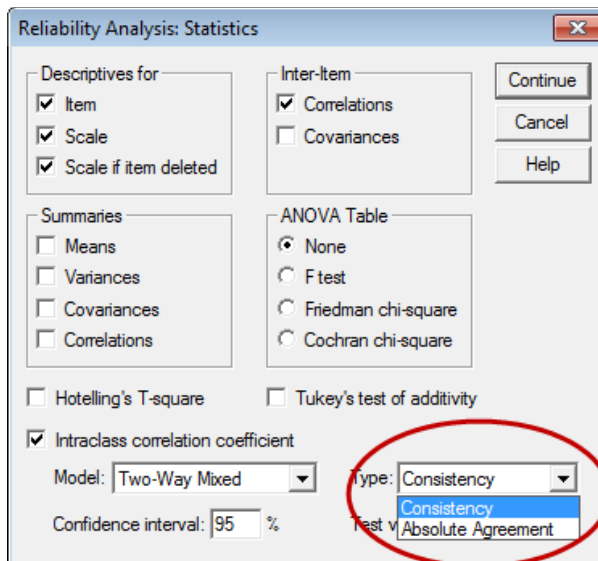
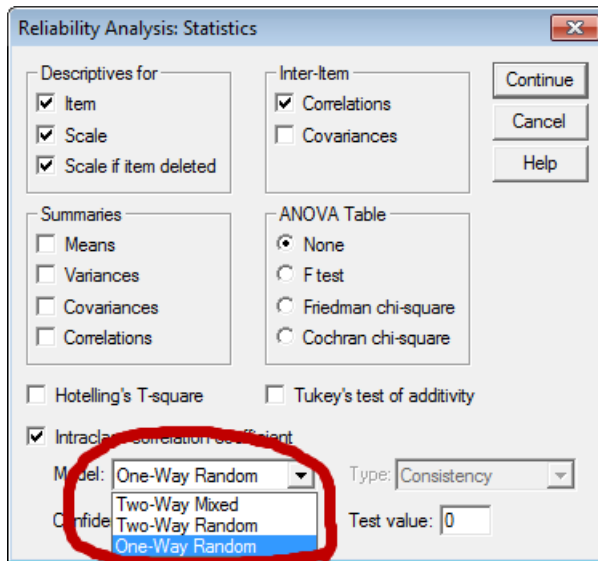
Average Measures = .775 --- This is the expected consistency if three judges provide scores and we take the mean of those three scores. Almost always better to have more than one judge when trying to obtain consistent scores from raters.

Use of one-way ANOVA for ICC is to be avoided if possible. Why? Because we lose information – use this approach only if we don't know which judges provide scores. Keep good records so we know which judges provided which scores. When we know which judges provided scores, we can then use Two-way ANOVA to obtain agreement measures for ICC rather than consistency measures as reported for the one-way ANOVA approach.

#### (b) Two-way ANOVA Random and Mixed Effects

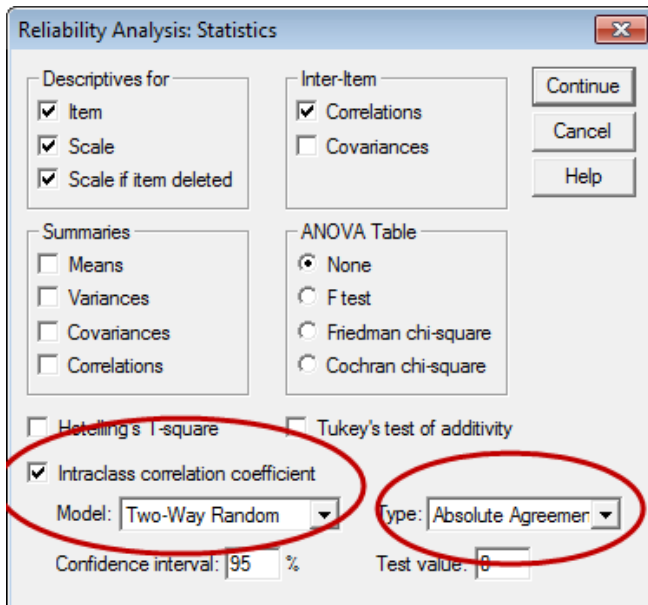
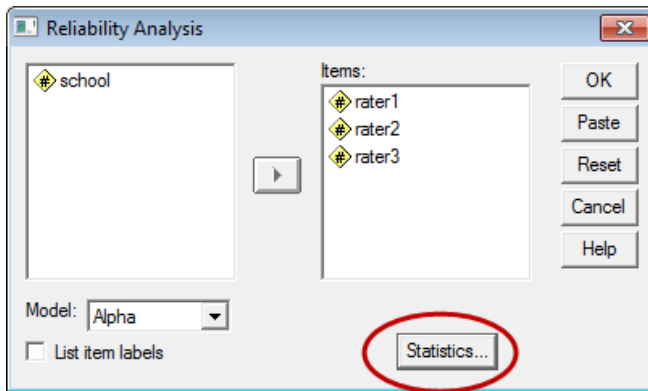
If one knows which raters provided which ratings, then use the two-way ANOVA option (see image below).

Estimates for the two-way ANOVA mixed and random are the same, so it does not matter which is selected. Also, choose Absolute Agreement since our interest is in whether raters provide similar ratings.



The difference between mixed and random two-way ANOVA depends upon how one views raters. If the raters are viewed as a random sample from a larger pool of raters, then use the random two-way ANOVA. If the raters are viewed as fixed and one is interested in inferences only for those raters, then use mixed.

Illustration of two-way ANOVA with specified raters.



Results are presented below.

Item Statistics

	Mean	Std. Deviation	N
rater1	5.0000	2.78887	10
rater2	7.0000	1.88562	10
rater3	5.2000	2.29976	10

Inter-Item Correlation Matrix

	rater1	rater2	rater3
rater1	1.000	.845	.641
rater2	.845	1.000	.564
rater3	.641	.564	1.000

The covariance matrix is calculated and used in the analysis.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
rater1	12.2000	13.733	.828	.754	.712
rater2	10.2000	21.289	.792	.715	.772
rater3	12.0000	20.222	.634	.413	.879

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
17.2000	38.622	6.21468	3

Intraclass Correlation Coefficient

	Intraclass Correlation <sup>a</sup>	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.558 <sup>b</sup>	.171	.847	6.870	9	18	.000
Average Measures	.791	.350	.944	6.870	9	18	.000

Two-way random effects model where both people effects and measures effects are random.

- Type A intraclass correlation coefficients using an absolute agreement definition.
- The estimator is the same, whether the interaction effect is present or not.

Single Measure = .558

If one plans to use a score from one rater, then the value of .558 indicates the level of consistent agreement one could expect – in this case it is too low to be judged consistent.

Average Measure = .791

If one plans to average scores from multiple raters, then the level of agreement is expected to be .791, which is much better than the value for the single rater. Generally taking several ratings and combining them into one overall rating is better – more precision and less error.

**x. Rater Agreement for Ordinal Data with Few Categories**

Material below to be updated

If ordinal data are used, some argue one should use Spearman rho or Kendall tau if there are only two judges or Kendall's coefficient of concordance if there are three or more.

Kendall's coefficient of concordance to be added.

ICC and Krippendorff's alpha

<http://www.john-uebersax.com/stat/icc.htm> (description of ICC and links to explaining SPSS implementation)

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4063345/>

Good table but for nominal data

[http://www.agreestat.com/book4/9780970806284\\_chap3.pdf](http://www.agreestat.com/book4/9780970806284_chap3.pdf)

41

**Table 14.13** Definitions of *ICC* with different models and notations used by different authors. The *ICC* measures with *k* in parentheses are defined for the average of *k* measurements, and the others are for single measurements.

ANOVA model	Interaction between rater and subject?	Authors		
		Shrout and Fleiss (1979)	McGraw and Wong (1996)	Barnhart et al. (2007)
One-way random effects		Case 1 $ICC(1,1)$ or $ICC(1,k)$	Case 1 $ICC(1)$ or $ICC(k)$	$ICC_1$
Two-way random effects	Without interaction	As below	Case 2A $ICC(A,1)$ or $ICC(A,k)$	$ICC_2$
	With interaction	Case 2 $ICC(2,1)$ or $ICC(2,k)$	Case 2 As above	$ICC_3$
Two-way mixed effects	Without interaction	As below	Case 3A $ICC(A,1)$ or $ICC(A,k)$	$ICC_2$
	With interaction	Case 3 $ICC(3,1)$ or $ICC(3,k)$	Case 3 As above	$ICC_3$

Regional Centre for Child and Youth Mental Health and Child Welfare

<http://folk.ntnu.no/slyderse/Pres24Jan2014.pdf>

- Barnhart et a. (2014). [Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting](#)
- Banerjee, Capozzoli, McSweeney, and Sinha (1999). [Beyond kappa: A review of interrater agreement measures.](#) Reviews a number of agreement measures.

## References

Gwet, K.L. (2014). Handbook of inter-rater reliability (4<sup>th</sup> ed.). Advanced Analytics, Gaithersburg, MD, USA.

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155-163.

Krippendorff, K. (2004). Reliability in Content Analysis: Some common Misconceptions and Recommendations. *Human Communication Research* 30,3: 411-433.

Liao, Hunt, & Chen (2010). Comparison between Inter-rater Reliability and Inter-rater Agreement in Performance Assessment. *Annals Academy of Medicine*. 613-618.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86, 420-428.

Tinsley H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology*, 22, 358-375.

Xie, Q. (2013). Agree or Disagree? A Demonstration of An Alternative Statistic to Cohen's Kappa for Measuring the Extent and Reliability of Agreement between Observers. Retrieved 9 April 2018 from:  
[https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/J4\\_Xie\\_2013FCSM.pdf](https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/J4_Xie_2013FCSM.pdf)