

Wood, J. M. (2007, October 3). Understanding and Computing Cohen's Kappa: A Tutorial. *WebPsychEmpiricist*. Retrieved October 3, 2007 from [http://wpe.info/papers\\_table.html](http://wpe.info/papers_table.html).

# WPE WebPsychEmpiricist

Understanding and Computing Cohen's Kappa: A Tutorial

October 3, 2007

James M. Wood

Department of Psychology, University of Texas at El Paso

Understanding and Computing Cohen's Kappa: A Tutorial

Cohen's Kappa (Cohen, 1960) is an index of interrater reliability that is commonly used to measure the level of agreement between two sets of dichotomous ratings or scores. This tutorial explains the underlying logic of Kappa and shows why it is superior to simple percentage of agreement as a measure of interrater reliability. Examples demonstrate how to calculate Kappa both by hand and with SPSS.

Correspondence regarding this article should be sent to: James Wood, Department of Psychology, University of Texas at El Paso, El Paso, TX 75220 or email to [jawood@utep.edu](mailto:jawood@utep.edu). This tutorial is copyrighted by James M. Wood but may be copied without permission for personal use and by instructors for classroom use.

## Understanding and Computing Cohen's Kappa: A Tutorial

### *Why Percentage Agreement is Not a Good Measure of Interrater Reliability*

A dichotomous rating involves a choice between two alternatives (e.g., yes or no, present or absent, pass or fail, accept or reject). If we want to know how well two raters agree when they make dichotomous ratings, there seems to be an obvious and straightforward approach to calculating agreement:

- (1) Count the total number of ratings made by each rater (Total Number) .
- (2) Count all of these ratings for which both raters agree  
(Number of Agreements).
- (3) Simply divide the Number of Agreements by the Total Number,  
to obtain "Percentage Agreement."

However, there is a subtle problem with this seemingly straightforward approach: It can make two raters appear to be highly reliable even if they are scoring completely at random! Consider two child psychologists who rated aggressive behavior in a child for twenty one-minute time periods. During each time period, each psychologist assigned a rating of "1" if the child acted aggressively and a rating of "0" if the child did NOT act aggressively. At the end of the twenty minutes, the ratings of the two psychologists looked like Table 1:

*Table 1: Ratings of Aggressive Behavior in a Child by Two Psychologists.*

Time Period	Ratings by Psychologist 1	Ratings by Psychologist 2
1	0	0
2	0	0
3	0	0
4	0	0
5	1	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	1
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0

As may be seen, the level of agreement between the two sets of ratings was very poor. Each psychologist indicated that the child acted aggressively during a single time period. However, Psychologist 1 indicated that the aggressive act occurred during Time Period 5, whereas Psychologist 2 indicated that it occurred during Time Period 12.

We can create a crosstabs table that compares the ratings of the two psychologists:

*Table 2: Crosstabs Table Comparing the Ratings by the Two Psychologists in Table 1.*

		Psychologist 1		Totals
		0	1	
Psychologist 2	0	18	1	19
	1	1	0	1
Totals		19	1	

As the table shows, in eighteen time periods (upper left cell) the two psychologists agreed that there was no aggressive behavior. In zero time periods (lower right cell) did both psychologists agree that aggressive behavior had occurred. There was one time period (upper right cell) in which Psychologist 1 thought that aggressive behavior had occurred, although Psychologist 2 did not. Conversely, there was one time period (lower left cell) in which Psychologist 2 thought that aggressive behavior had occurred, although Psychologist 2 did not.

It is instructive to calculate Percentage Agreement for this example. The Total Number of ratings is 20. The Number of Agreements is 18 (i.e. 18 time periods when both psychologists agreed there was NO aggressive behavior, and 0 time periods when both psychologists agreed that aggressive behavior had occurred). Percentage agreement can be calculated by dividing the Number of Agreements by the Total Number of ratings:

$$\text{Percentage agreement} = 18 / 20 = 90\%$$

As may be seen, the level of percentage agreement looks very good (90%), even though the two psychologists can't seem to agree at all about when and if aggressive behavior has occurred. How is it possible that two raters can achieve 90% agreement if they are scoring randomly?

At first thought, it might seem that two scorers who are scoring randomly should achieve only 50% agreement, and that 90% is substantially above agreement. However, as the example

shows, "chance" agreement between two raters can actually be much higher than 50%. Let us consider the example more closely. Psychologist 1 and Psychologist 2 each score aggression as "present" in .05 (1/20) of the one-minute periods. Therefore, even if they are scoring randomly, they will occasionally agree by chance alone that aggression is present. The exact proportion of expected agreements for "aggression present" can be calculated as follows:

$$.05 \times .05 = 1/20 \times 1/20 = .0025$$

Similarly, Psychologist 1 and Psychologist 2 each score aggression as "absent" in .95 (19/20) of the one-minute periods. Therefore, even if they are scoring randomly, they will very frequently agree by chance alone that aggression is absent. The exact proportion of expected agreements for "aggression absent" can be calculated as follows:

$$.95 \times .95 = 19/20 \times 19/20 = .9025$$

The total proportion of expected agreements (E) can then be calculated by adding together the expected agreements for "aggression present" and "aggression absent":

$$E = .9025 + .0025 = .9050$$

As may be seen, the proportion of agreement expected by chance alone (.9050) is almost exactly the proportion of agreement that was actually observed (.90) between Psychologist 1 and Psychologist 2. In other words, the numbers show that the agreement between the two psychologists was approximately what would be expected by chance.

Because percentage agreement can make even random ratings look "good," it has been thoroughly and repeatedly criticized by statisticians since the 1960s. There is now nearly unanimous agreement among statisticians that a statistic known as Cohen's Kappa is a much better measure of interrater reliability than Percentage Agreement is.

### ***Description of Kappa***

Kappa (Cohen, 1960; Fleiss, Levin, & Paik, 2003; Spitzer, Cohen, & Fleiss, 1968) is very closely related to the familiar correlation coefficient, Pearson's  $r$ . For example, kappa and Pearson's  $r$  are usually very close (within .05 of each other) if they are calculated for the same set of dichotomous ratings from two raters. Like Pearson's  $r$ , Kappa can range anywhere from -1.0 to +1.0. A kappa of 1.0 means that two raters show perfect agreement, a kappa of -1.0 means that they show perfect and consistent disagreement, and a kappa of 0 means that the two raters show a random level of agreement/disagreement (i.e. there is no relationship between their ratings).

What level of interrater reliability is "good enough"? For research purposes, there seems to be general agreement that the kappa should be at least .60 or .70. However, if the ratings or decisions are to be used for making applied decisions about a particular individual (e.g. scoring of intelligence tests), reliability should probably be higher, so that kappa is at least .80 or .90. In addition, it is worth noting that psychiatric diagnoses and general medical diagnoses generally range from a kappa of .40 to a kappa of .90, with an average reliability between .60 and .70. It seems that adequate reliability is easier to obtain for individual test scores or narrowly limited behavioral ratings than for diagnoses

### ***Calculation of Kappa by Hand***

Kappa for a set of ratings can easily be calculated with SPSS software, as will be explained in detail in a later section of this article. However to really understand Kappa and its logic, it is a good idea to calculate it by hand a few times. The formula for Kappa is:

(Observed percentage of agreement) - (Expected percentage of agreement)

-----

1 - (Expected percentage of agreement)

or in abbreviated form

$$\text{Kappa} = \frac{\text{O} - \text{E}}{1 - \text{E}}$$

The "observed percentage of agreement" is the proportion of ratings where the scorers are in agreement.

The "expected percentage of agreement" is the proportion of agreements that would be expected "by chance" between the raters if they were scoring randomly.

So conceptually, kappa is equal to the proportion of agreement actually observed between raters, after adjusting for the proportion of agreement expected "by chance" (randomly) For this reason, some writers refer to Kappa as the "chance-corrected proportion of agreement."

The steps in calculating kappa are as follows:

Step 1. Construct a crosstabs table of the data (Table of Raw Frequencies)

Step 2. From the Raw Frequencies table, construct a second crosstabs table that contains proportions rather than raw frequencies (Table of Proportions).

Calculate O = the sum of the upper left cell and the lower right cell.

Step 3. From the Table of Proportions, calculate the "marginals," which are the sums of the cells in each row and column.

Step 4. Construct a third crosstabs table.

(a) In the upper left cell insert the following product:

Marginal of the top row X marginal for the left column (using the marginals from the Table of Proportions that you constructed in Step 3)..

(b) In the lower right cell insert the following product:

Marginal of the bottom row X marginal for the right column.

(c) Calculate E = the sum of the upper left cell and the lower right cell.

Step 5. Insert these values into the following formula

$$\text{Kappa} = \frac{\text{O} - \text{E}}{1 - \text{E}}$$

### ***Calculating Kappa By Hand: An Example***

This section shows how kappa can be calculated by hand following the steps just described.

Some readers may prefer to skip this section and continue to the next section, “Understanding the Formula for Kappa.”

### **Example (based on the data in Table 2):**

Step 1: Construct a table of raw frequencies.

		Psychologist 1		
		0	1	
Psychologist 2	0	18	1	19
	1	1	0	1
	Totals	19	1	

Step 2: Convert the table in Step 1 into frequencies and then calculate O (Observed Percentage of Agreement)

		Psychologist 1	
		0	1
Psychologist 2	0	.90	.05
	1	.05	.00

$$O = .90 + .00 = .90$$

Step 3: Calculate the marginals for the table in Step 2.

		Psychologist 1		Totals
		0	1	
Psychologist 2	0	.90	.05	.95
	1	.05	.00	.05
Totals		.95	.05	

Step 4: Cross-multiply appropriate marginals to arrive at cross-products, then sum these cross-products to arrive at E (Expected Percentage of Agreement).

		Psychologist 1		Totals
		0	1	
Psychologist 2	0	.9025		← .95
	1		.0025	← .05
Totals		↑ .95	↑ .05	

$$E = .9025 + .0025 = .9050$$

Step 5: Plug the values from Steps 2 and 4 into the formula for kappa.

$$\text{Kappa} = \frac{O - E}{1 - E} = \frac{.90 - .9050}{1 - .9050} = \frac{-(.0050)}{.095} = -.0526$$

### ***Understanding the Formula for Kappa***

Kappa represents (approximately) the correlation between the scores of two raters. It is very closely related to a statistic called the phi coefficient, which in turn is very closely related to Pearson's  $r$ .

The formula for Kappa is very simple and relatively easy to understand. You will recall that  $E$  = the percentage of agreement between two raters that would be expected "by chance" (i.e. if the two raters are scoring randomly). For instance, in the example given earlier, the proportion of agreement expected by chance between Psychologist 1 and Psychologist 2 was .9050. The *maximum* agreement possible between any two raters (if they are in perfect agreement) is 1.00.

Therefore, the maximum amount that Psychologist 1 and Psychologist 2 could improve above chance agreement is calculated as follows:

$$\begin{aligned} \text{Perfect Agreement} - \text{Chance Agreement} &= 1 - E \\ &= 1.00 - .9050 \\ &= .0950 \end{aligned}$$

As may be seen, the denominator of Kappa ( $1 - E$ ) simply represents the *maximum amount that the two raters could have improved* above chance agreement.

Similarly, the numerator of Kappa equals the percentage of agreement that was *actually observed* between the raters ( $O$ ) minus the percentage of agreement *expected by chance* ( $E$ ). Thus for Psychologist 1 and Psychologist 2:

$$\begin{aligned} \text{Observed Agreement} - \text{Chance Agreement} &= O - E \\ &= .90 - .9050 \\ &= -.0050 \end{aligned}$$

As may be seen, the numerator of Kappa ( $O - E$ ) simply represents the amount the raters *actually did improve* beyond chance agreement.

When we put the numerator and the denominator together, then, we now understand Kappa in a new way: Kappa is equal to the amount of improvement that raters *actually showed* above chance, divided by the maximum amount of improvement that they *could have shown*.

$$\text{Kappa} = \frac{O - E}{1 - E} = \frac{-(.0050)}{.0950} = -.0526$$

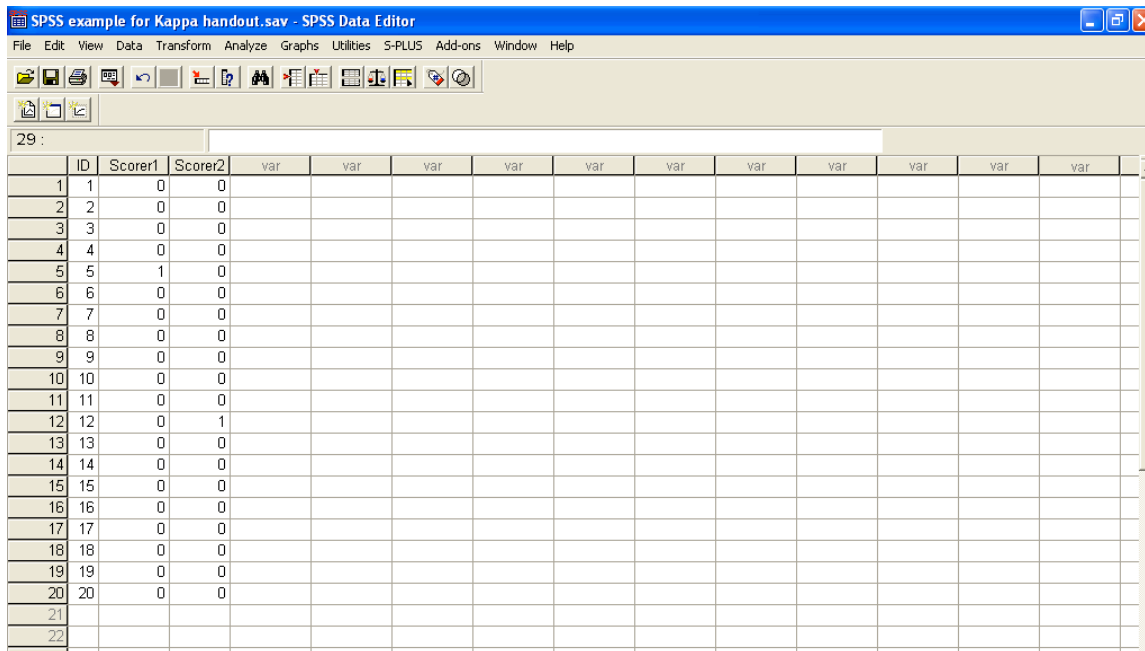
As can be seen Kappa in this example is very close to zero. In other words, Kappa shows (as we intuitively already know) that the level of agreement between the two psychologists was approximately random.

### *Calculation of Kappa Using SPSS*

Calculating Kappa by hand a few times is a good learning exercise. However, when analyzing real data, it's more convenient to do the calculations with SPSS. This section explains how.

First the data should be entered into an SPSS data sheet, as shown below.

As can be seen, the scoring data from the two psychologists is entered as two variables, "Scorer1" and "Scorer2". Each row of data represents the scores of aggression for a particular time period.

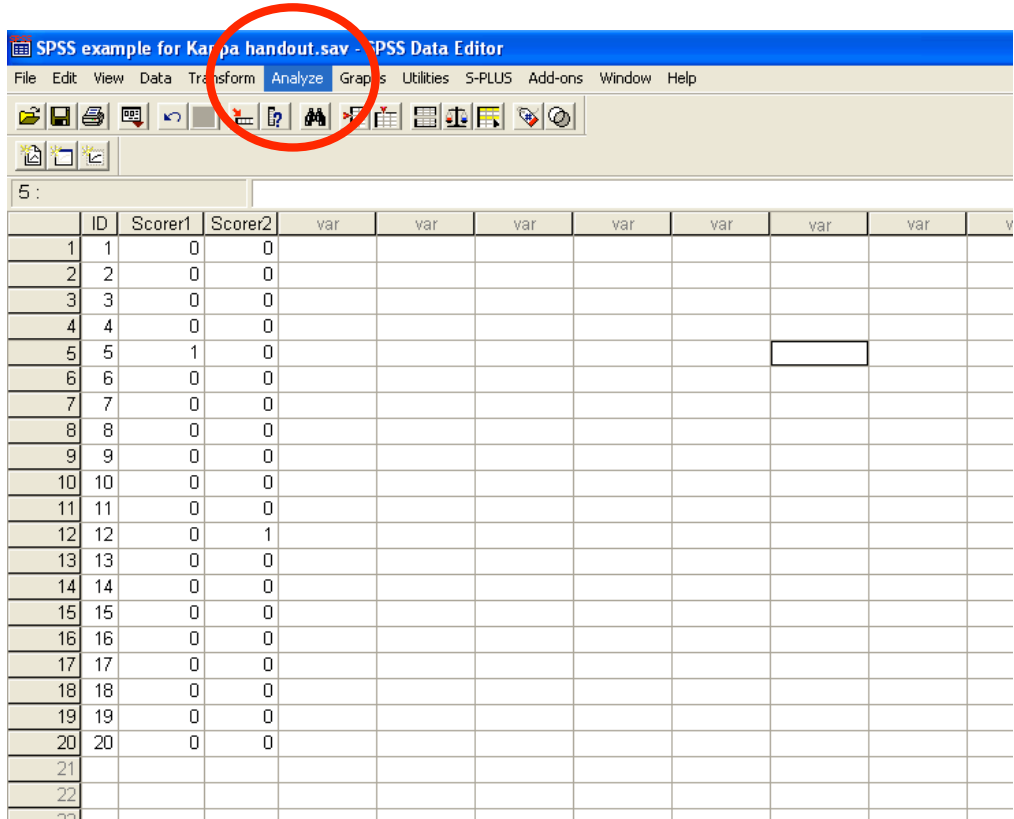


SPSS example for Kappa handout.sav - SPSS Data Editor

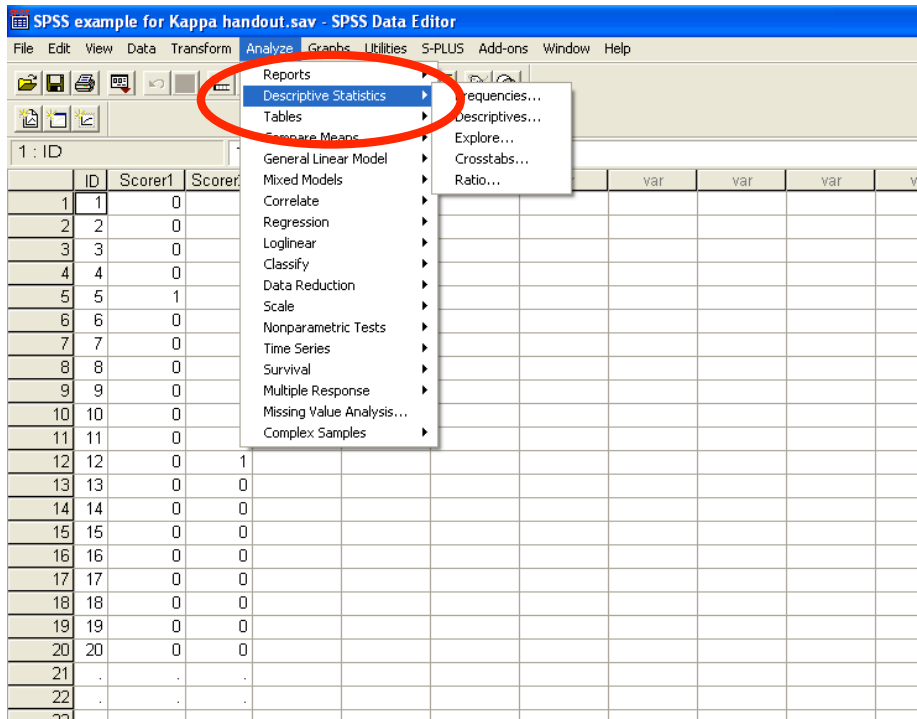
File Edit View Data Transform Analyze Graphs Utilities S-PLUS Add-ons Window Help

	ID	Scorer1	Scorer2	var	var	var	var	var	var	var	var	var	var	var	var
1	1	0	0												
2	2	0	0												
3	3	0	0												
4	4	0	0												
5	5	1	0												
6	6	0	0												
7	7	0	0												
8	8	0	0												
9	9	0	0												
10	10	0	0												
11	11	0	0												
12	12	0	1												
13	13	0	0												
14	14	0	0												
15	15	0	0												
16	16	0	0												
17	17	0	0												
18	18	0	0												
19	19	0	0												
20	20	0	0												
21															
22															

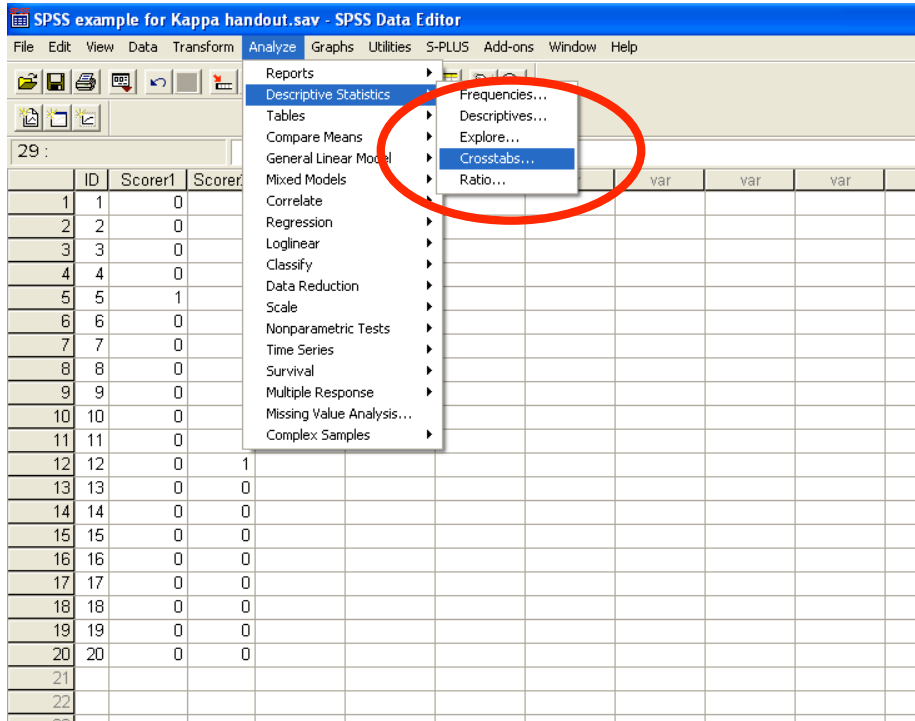
After entering the data, click the "Analyze" menu.



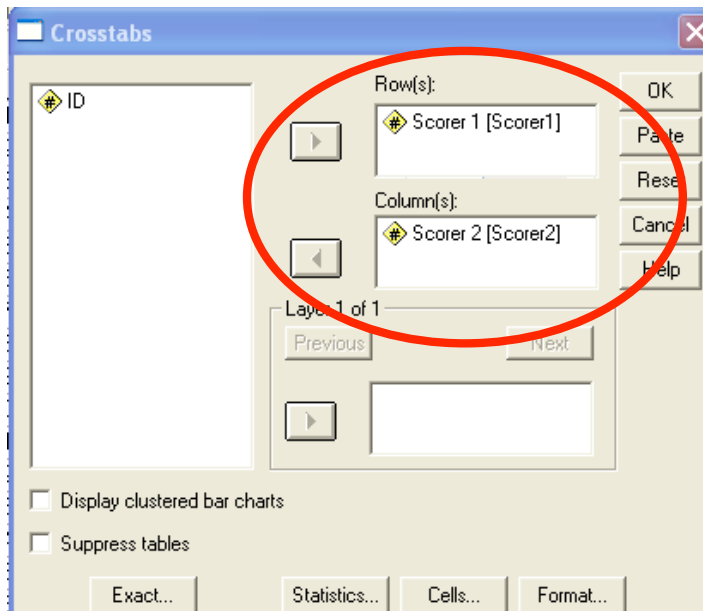
Next click "Descriptive Statistics."



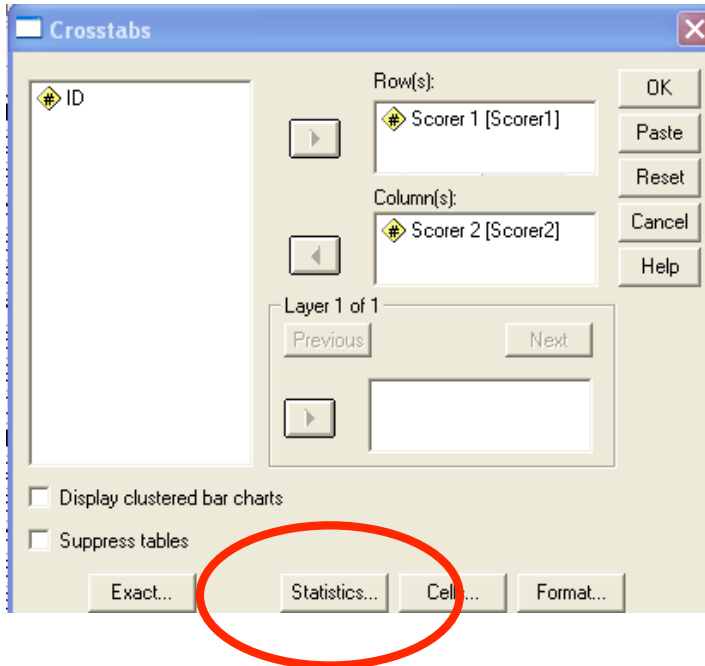
Next click "Cross-tabs."



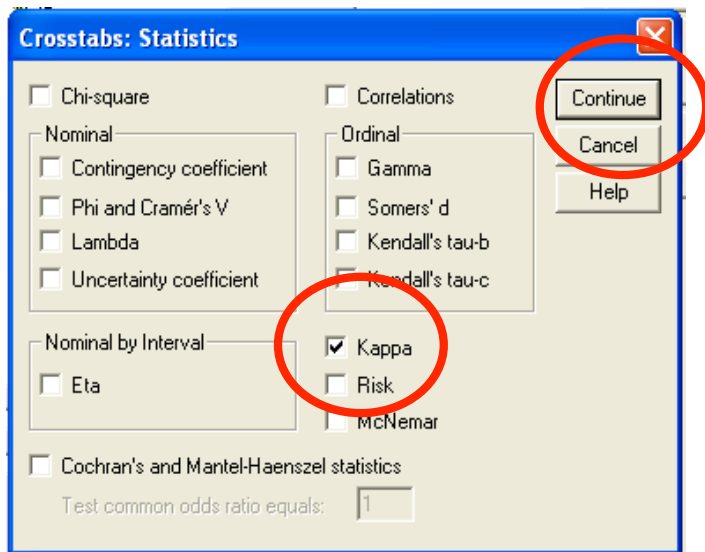
A crosstabs window will now open. Use the arrows to select Scorer 1 for "Row(s)" and Scorer 2 for "Column(s);"



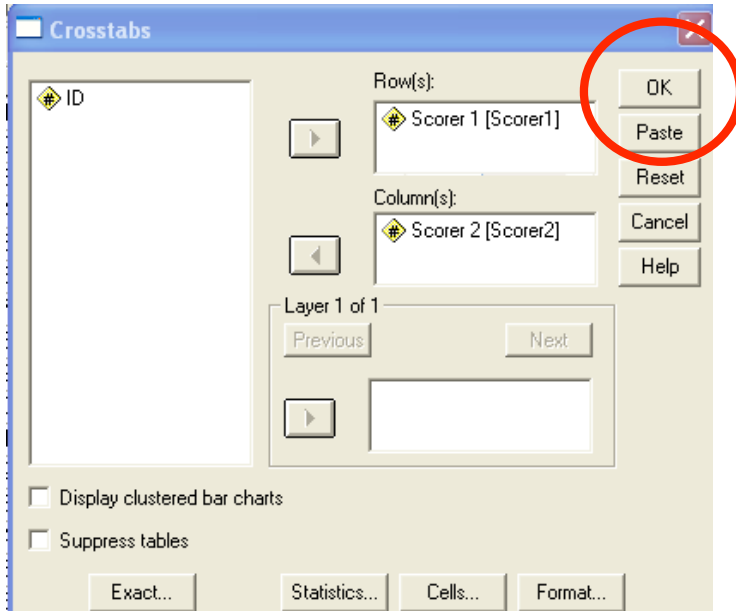
Next click the "Statistics" Option



A window will open for "Crosstabs. Check the box for "Kappa," then click Continue.



Finally, in the Crosstabs window, click "OK."



The SPSS output for the two psychologists' scores is shown below. It's encouraging to notice that SPSS arrived at the same value of Kappa (-.053) as the hand calculations did, except for rounding.

**Crosstabs**

**Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Scorer 1 * Scorer 2	20	50.0%	20	50.0%	40	100.0%

**Scorer 1 ' Scorer 2 Crosstabulation**

Count

		Scorer 2		Total
		0	1	
Scorer 1	0	18	1	19
	1	1	0	1
Total		19	1	20

**Symmetric Measures**

		Value	Asymp. Std. Error <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
→ Measure of Agreement	Kappa	-0.053	.037	-.235	.814
	N of Valid Cases	20			

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

Although the printout reports the statistical significance of Kappa (.814), this number is not generally of much interest. In most studies of interrater reliability, the important question is not "Is Kappa significantly different from zero?" but rather "Is Kappa high enough to meet the standards for use in research or in clinical settings?" In the present case, Kappa is much lower than would be considered acceptable for either research or clinical work.

This concludes the tutorial on Kappa. I hope it has been helpful to you.

### ***References***

- Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Fleiss, J. L., Levin, B., & Paik M. C. (2003). *Statistical methods for rates and proportions*. Hoboken, New Jersey: Wiley.
- Spitzer, R. L., Cohen, J. A., & Fleiss, J. L. (1968). Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry, 17*, 83-87.