Cohen’s Kappa in SPSS: A Comprehensive Guide to Inter-Rater Reliability

Authored by

Mohammed looti

November 12, 2025

Introducing Cohen's Kappa: Assessing Reliability Beyond Chance

Cohen's Kappa is an indispensable statistical measure specifically designed to quantify the degree of agreement between two independent observers, often referred to as **raters**, when they categorize items into distinct, mutually exclusive categories. While a simple calculation of percentage agreement might initially seem sufficient, it often produces misleading results because it fails to account for the agreement that is expected to occur purely by random chance. Kappa addresses this fundamental flaw, providing a coefficient that truly reflects the inter-rater reliability inherent in the rating process.

The primary goal of employing Cohen's Kappa is to establish a robust level of consistency in judgment across various fields. Whether applied in **medical diagnostics**, psychological assessment, or quality control, this statistic ensures that classification outcomes are reliable, regardless of which rater performs the evaluation. For example, if two professionals classify 100 observations, a high observed agreement (say, 90%) might be skewed if the categories themselves are heavily imbalanced. Kappa corrects for this inherent statistical bias, yielding a coefficient that isolates the true reliability attributable to the quality of the rating instrument and the raters' skill.

It is vital to recognize the specific scope of this statistic: Kappa is optimally applied when dealing with nominal or categorical data and involves precisely two raters. Utilizing Kappa guarantees that the conclusions drawn from the categorical data--such as identifying patterns in qualitative research or assessing product defects--are grounded in a reliable foundation, thereby significantly enhancing the validity and trustworthiness of the overall study findings.

Decoding the Kappa Formula and Interpreting the Coefficient

The mathematical elegance of Cohen's Kappa lies in its ability to isolate the true agreement by subtracting the chance agreement from the observed agreement. The formula calculates the proportion of times the raters agree beyond the hypothetical probability that they would concur randomly. This critical adjustment distinguishes Kappa as a superior measure compared to rudimentary percentage calculations.

The formula for Cohen's Kappa is conventionally expressed as:

k = (po - pe) / (1 - pe)

The components of this equation are defined as follows, reflecting the necessary elements for statistical calculation:

po: Represents the relative **observed agreement** among the raters. This is the proportion of

observations where both raters assigned the identical category.
pe: Stands for the hypothetical **probability of chance agreement**. This value is derived from the marginal totals of each rater's classification distribution, estimating how often they would agree randomly based on category prevalence.

The resulting Kappa coefficient is typically scaled between **0** and **1**. A coefficient of **0** indicates that the observed agreement is no better than what would be expected purely by chance, suggesting a complete lack of genuine inter-rater reliability. Conversely, a value of **1** signifies **perfect agreement**, meaning the raters' classifications matched for every single item. Although rare, a negative Kappa value is possible, implying systematic disagreement--agreement that is worse than random.

To standardize reporting, researchers rely on established benchmarks for interpreting the numerical Kappa output. These qualitative guidelines transform the coefficient into actionable insight regarding the strength of the agreement. The image below summarizes the commonly accepted standards used across various statistical disciplines:

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Structuring Data for Kappa Calculation: An Expert Judgment Scenario

To demonstrate the practical steps required to calculate Cohen's Kappa, we will utilize a compelling example involving expert judgment. Imagine a scenario where two seasoned art museum curators are tasked with evaluating a collection of 30 paintings. Their mandated task is to classify each piece into one of two **mutually exclusive categories**: "Good Enough" (Yes) or "Not Good Enough" (No) for potential inclusion in a major exhibition. This binary, two-rater structure makes the dataset perfectly suited for Kappa analysis.

The required dataset structure is specific: it must contain two variables, one for each curator (Rater 1 and Rater 2), where every row represents a single, paired observation (i.e., one painting rated by

both experts). The data must clearly reflect the categorical outcome assigned by each individual.

The raw ratings provided by the two curators for the 30 paintings are displayed below. This structure, where the columns are the raters' judgments and the rows are the cases (paintings), is the mandatory format for inputting data into IBM SPSS Statistics for subsequent cross-tabulation:

	Rater1	Rater2	var	var	var	va
1	Yes	Yes				
2	Yes	Yes				
3	Yes	Yes				
4	Yes	Yes				
5	Yes	Yes				
6	Yes	No				
7	Yes	Yes				
8	Yes	No				
9	Yes	Yes				
10	Yes	No				
11	Yes	No				
12	Yes	Yes				
13	Yes	Yes				
14	Yes	Yes				
15	Yes	Yes				
16	Yes	No				
17	Yes	Yes				
18	Yes	Yes				
19	Yes	No				
20	No	Yes				
21	No	Yes				
22	No	No				
23	No	No				
24	No	Yes				
25	No	Yes				
26	No	No				
27	No	No				
28	No	No				
29	No	No				
30	No	No				
31						

Our objective is to formally quantify the degree of consistency between Curator 1 and Curator 2. While a simple agreement percentage provides a superficial measure, the calculation of Cohen's

Kappa will yield the statistically adjusted coefficient that rigorously corrects for chance agreement, offering a much more accurate and reliable measure of their inter-rater reliability.

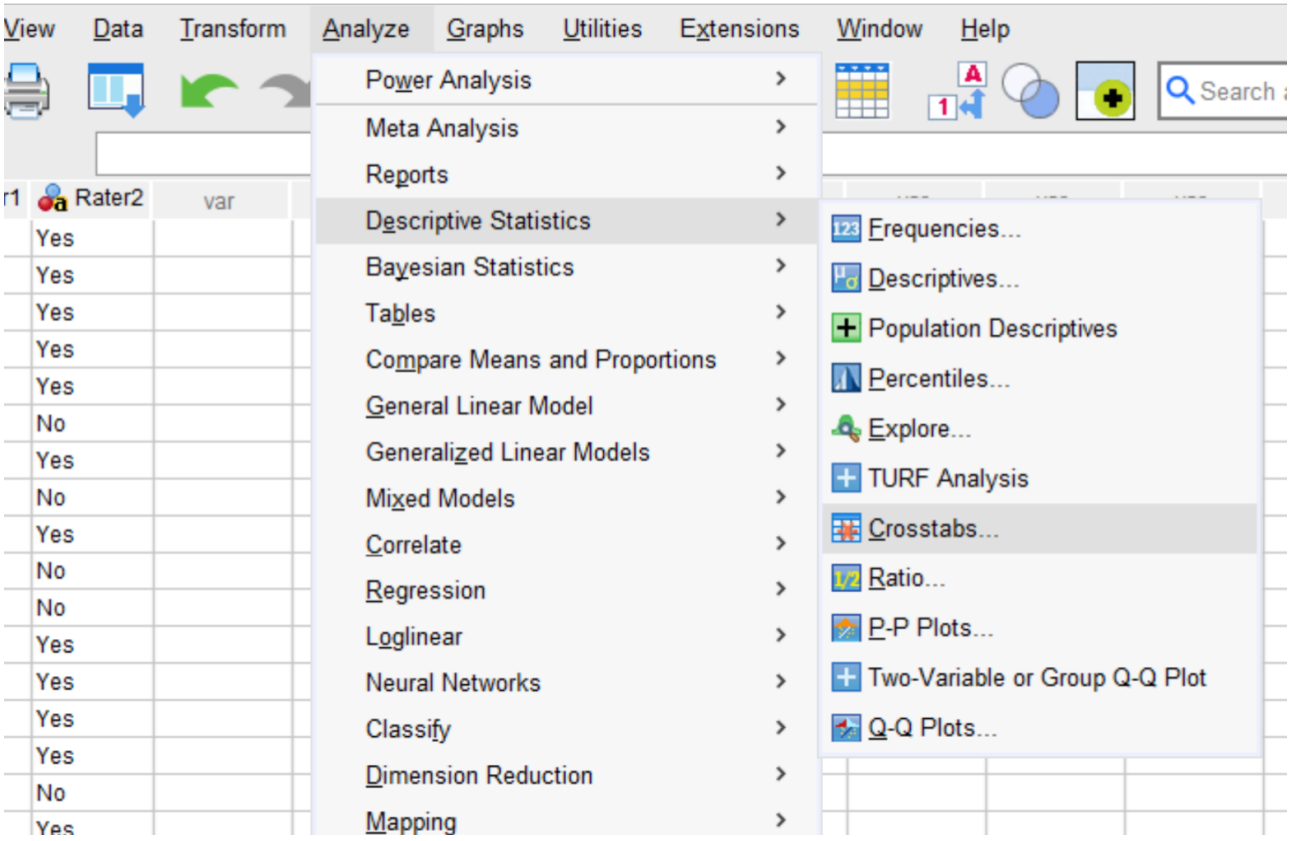
Executing the Analysis: Step-by-Step Guide in SPSS

Calculating Cohen's Kappa within the **IBM SPSS Statistics** environment is streamlined using the Crosstabs functionality, a tool specifically designed to analyze the relationship between two categorical variables. This process begins by navigating the software's main analysis menu to locate the appropriate statistical procedure.

To initiate the calculation, follow this precise sequence in the SPSS menu bar:

Click the **Analyze** tab located at the top of the interface.
Hover over **Descriptive Statistics** in the dropdown menu.
Select **Crosstabs** to open the main dialog box.

This path directs the software to the necessary tool for generating the contingency table, which is the foundation for calculating the Kappa coefficient and other associated measures of association.



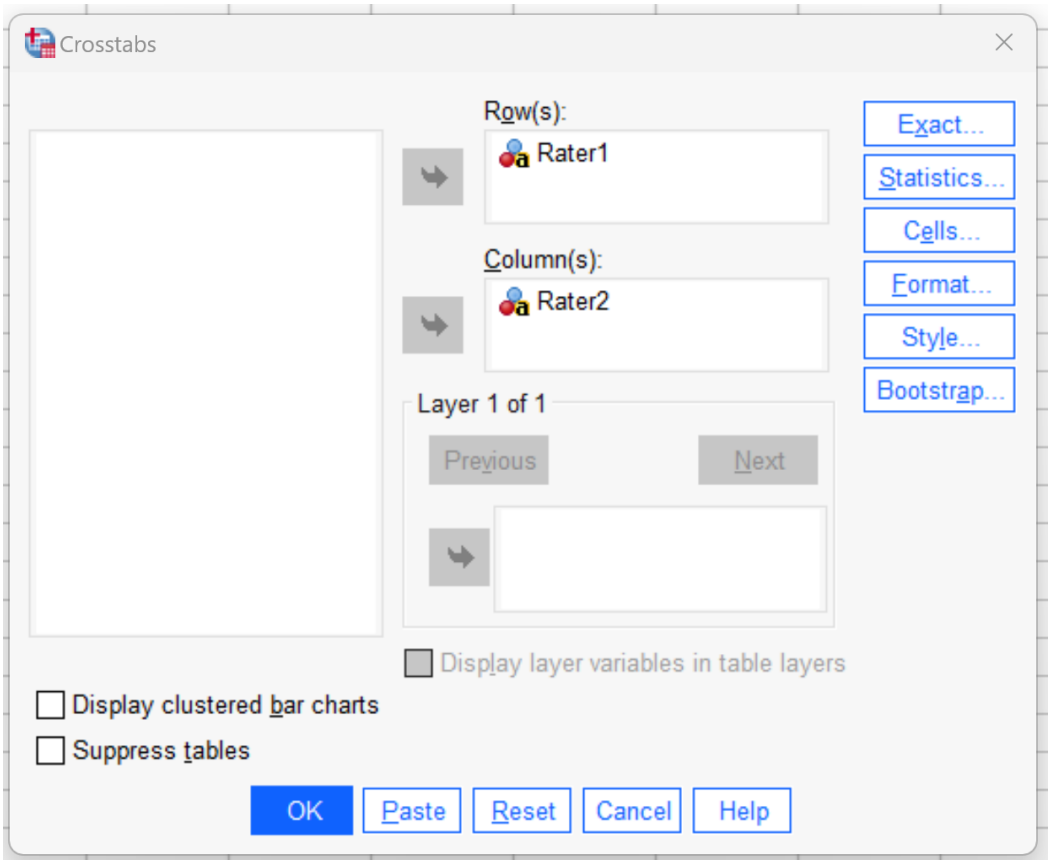
Once the Crosstabs dialog box is displayed, the two rater variables must be correctly mapped to the row and column dimensions. Although the assignment of which rater goes into the **Row** versus

the **Column** field does not impact the final Kappa value, standard practice dictates placing one variable in each dimension for clear interpretation of the resulting matrix.

Configure the variables within the Crosstabs window as follows:

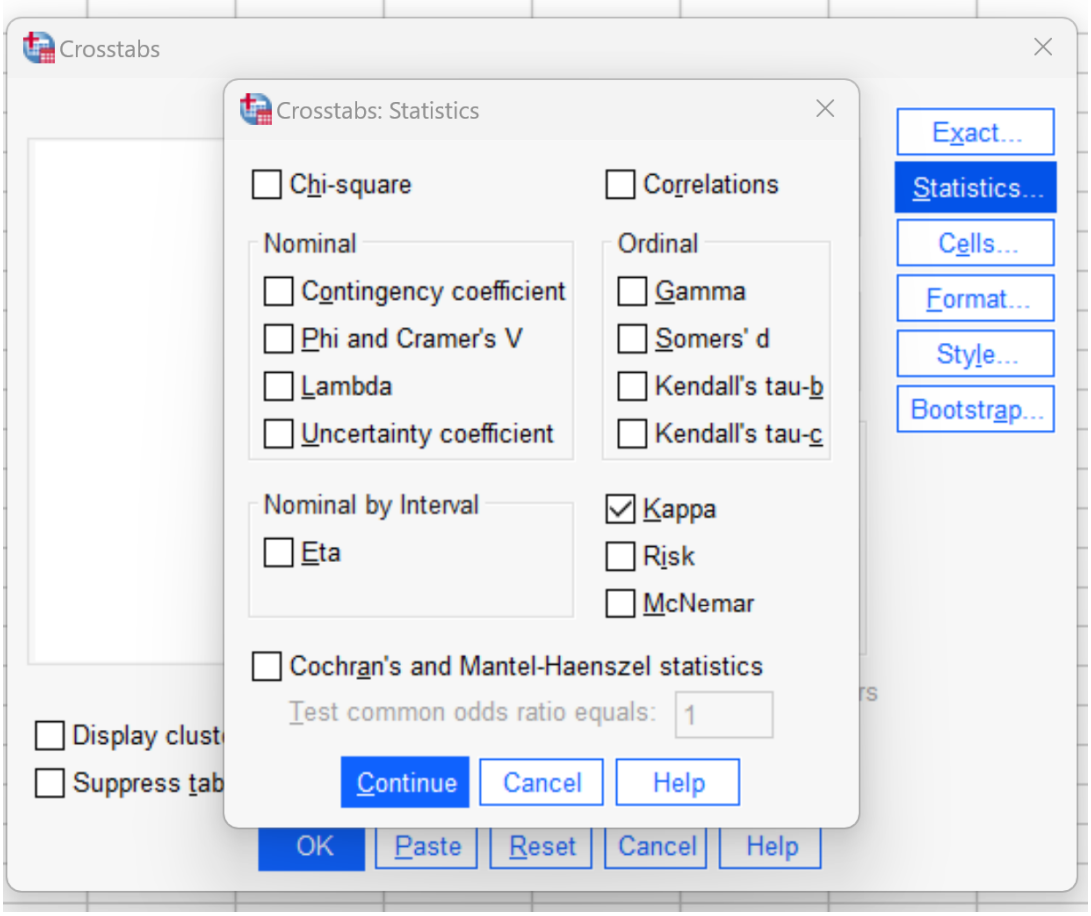
Drag the variable representing **Rater1** into the **Rows** box.
Drag the variable representing **Rater2** into the **Columns** box.

This configuration establishes the necessary two-by-two matrix structure that SPSS will use to compute the frequency counts of agreement and disagreement.



The final critical step before running the analysis is explicitly instructing SPSS to calculate the Kappa statistic. This is achieved through the **Statistics** sub-dialogue. Click the **Statistics** button, which opens a dedicated panel for specifying non-default measures of association and agreement.

In the Statistics window, locate and check the box labeled **Kappa**. This selection ensures that the Cohen's Kappa coefficient will be generated alongside the standard crosstabulation table. After confirming the selection, click **Continue** to close the Statistics window, and then click **OK** in the main Crosstabs dialog box to execute the analysis and produce the statistical output.



Analyzing the SPSS Output: Crosstabulation and the Kappa Coefficient

Upon successful execution, **SPSS** generates several output tables crucial for interpreting the results. The first key component is the **Crosstabulation table**, which provides a detailed frequency summary, outlining every combination of ratings produced by the two curators. This table is indispensable for understanding the raw agreement patterns before any adjustment for chance is applied.

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Rater1 * Rater2	30	100.0%	0	0.0%	30	100.0%

Rater1 * Rater2 Crosstabulation

Count

		Rater2		
		No	Yes	Total
Rater1	No	7	4	11
	Yes	6	13	19
Total		13	17	30

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.309	.175	1.708	.088
N of Valid Cases		30			

a. Not assuming the null hypothesis.

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

A careful examination of the crosstab reveals the specific counts of agreement and disagreement in our example:

Agreement on "No": The cell where Rater 1 (Row) and Rater 2 (Column) both assigned "No" shows **7** paintings.
Agreement on "Yes": The cell where both Raters assigned "Yes" shows **13** paintings.
Disagreement (Rater 1 No, Rater 2 Yes): There were **4** instances of disagreement here.
Disagreement (Rater 1 Yes, Rater 2 No): This category contained **6** paintings.

The total count of agreed-upon ratings is 7 + 13 = 20, out of a total sample size of 30 paintings. This results in a simple observed agreement of approximately 66.7%. However, as previously emphasized, this percentage is inflated by chance agreement.

The definitive result is presented in the final output table, typically labeled "Symmetric Measures." In this specific analysis, the calculated **Cohen's Kappa coefficient** is found to be **.309**. This figure

represents the true, chance-corrected measure of agreement between the two art curators. Crucially, **SPSS** also provides the asymptotic standard error and the significance level (p-value), which allows researchers to test the null hypothesis that the agreement between the raters is zero.

Interpreting the Coefficient: Assessing the Level of Agreement

The numerical output of **.309** must now be translated into a qualitative assessment of the inter-rater reliability using established statistical guidelines. This interpretive step is essential for converting a raw statistical output into a meaningful conclusion regarding the consistency of the curators' judgments.

Referring back to the standardized interpretation table, we contextualize our calculated value:

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Based on this widely accepted scale, a coefficient of **.309** falls squarely within the range typically defined as having a **"Fair" level of agreement**. This result clearly indicates that while the curators agreed on a majority of the paintings (66.7% observed agreement), a statistically significant portion of that concurrence is likely attributable to random chance rather than perfect consistency in applying the rating criteria. Therefore, their true reliability, when adjusted for random likelihood, is moderate at best.

A finding of "Fair" agreement carries important implications for the study methodology. It suggests potential weaknesses that require immediate attention, such as ambiguous rating criteria, inadequate training for the raters, or a lack of clarity regarding how to handle borderline cases. If the research required **substantial** or **near-perfect agreement** (as is often the standard in sensitive fields like clinical diagnosis), a Kappa of .309 would be deemed insufficient. The true value of Cohen's Kappa is its ability to highlight these inconsistencies, guiding researchers to refine their data collection instruments and ultimately leading to more reliable and valid conclusions.

Further Exploration in Reliability Statistics

For analysts and researchers seeking to deepen their understanding of reliability statistics and explore methodologies beyond the two-rater scenario, several related measures are available.

Detailed tutorials comparing **Cohen's Kappa** with **Fleiss' Kappa** (used when more than two raters are involved).
Comprehensive explanations detailing the application and interpretation of **weighted** versus **unweighted Kappa statistics**, which account for the magnitude of disagreement.