How to calculate Fleiss’ Kappa in Excel?

Authored by

stats writer

December 24, 2025

The reliability of measurement is a cornerstone of rigorous research, especially in fields like psychology, market research, and medical diagnostics, where human judgment plays a critical role. When multiple assessors or judges--known as raters--are tasked with classifying items or subjects into specific categories, it is imperative to quantify the consistency of their classifications. This necessity leads us to the critical measure known as Fleiss' Kappa, a powerful statistic designed to assess the degree of reliability across more than two raters. Unlike simpler measures that might only look at raw percentage agreement, Kappa statistics adjust for the agreement that would be expected to occur purely by chance, offering a much more robust measure of consensus. Understanding how to calculate and interpret this metric is essential for validating the quality of data collected through observational or subjective rating methods. Although specialized statistical software is often preferred for complex calculations, Microsoft Excel remains a highly accessible and practical tool for performing this analysis, provided the necessary steps are meticulously followed.

Initially, it is worth noting the theoretical background of Fleiss' Kappa. It extends the logic of Cohen's Kappa, which is limited to two raters, allowing researchers to evaluate consensus among any fixed number of raters (M), provided they rate a set of N items into K fixed categorical ratings. The resulting Kappa value provides a standardized measure of inter-rater agreement, indicating the proportion of agreement observed beyond what chance alone would predict. A crucial initial step in any Kappa calculation involves structuring the data precisely. This structure typically requires a matrix where rows represent the subjects or items being rated, and columns represent the categories, with the cell values containing the count of raters who assigned that item to that specific category. This foundational setup ensures that all subsequent calculations, which involve determining proportions and expected agreements, are accurate and reflective of the underlying rating behavior.

The range of Fleiss' Kappa provides an immediate, intuitive interpretation of the reliability achieved. The metric is constrained between 0 and 1, though negative values are theoretically possible if the observed agreement is worse than random chance, suggesting systemic misunderstanding or bias among the raters. A value of 0 signifies that the observed agreement is no better than what would be achieved if the raters were assigning categories randomly, meaning there is no detectable consensus above chance level. Conversely, a value of 1 represents perfect agreement, where every single rater assigns the same category to every single item, indicating maximum possible consistency and reliability. Most real-world scenarios yield a Kappa value somewhere between these two extremes. Interpreting these intermediate values often requires reference to established guidelines, typically adapted from scales used for Cohen's Kappa, which help researchers classify the strength of the agreement as poor, fair, moderate, good, or very good, thereby contextualizing the utility of the collected rating data.

Defining and Understanding Fleiss' Kappa

Fleiss' Kappa, often denoted by $kappa$, serves a vital purpose in establishing the objectivity of human-assigned data. When research depends on subjective classification--such as diagnosing diseases, grading essays, or evaluating product quality--the extent to which different experts arrive at the same conclusion determines the trustworthiness of the entire dataset. Without demonstrable high inter-rater agreement, the resulting data are highly susceptible to individual rater bias, rendering statistical conclusions unreliable. Therefore, calculating this statistic moves beyond simple descriptive percentages; it provides a stringent test of reliability by factoring in the element of randomness. The formal calculation compares the observed agreement (the actual consistency seen in the ratings) with the agreement expected by chance, standardizing this difference against the maximum possible agreement beyond chance. This rigorous methodology is why Kappa is considered superior to crude agreement percentages, especially when categories are unbalanced or when the number of available choices is small.

To fully appreciate the metric, one must understand the components it incorporates. The formula relies fundamentally on the proportions of raters agreeing on each category for each item, and the overall distribution of categories across all items and raters. Specifically, it involves calculating p_j, which is the marginal proportion of all ratings assigned to category j, and P_i, the proportion of agreement for item i. These intermediate steps are crucial because they inform the calculation of the two main structural elements of the formula: P_o (the mean observed proportion of agreement) and P_e (the mean expected proportion of agreement). The observed agreement, P_o, represents how often raters truly aligned, averaged across all items. The expected agreement, P_e, represents the hypothetical scenario where raters assign categories based solely on the marginal probabilities of those categories occurring, without any genuine consensus mechanism. The final Fleiss' Kappa formula is essentially $kappa = (P_o - P_e) / (1 - P_e)$, elegantly capturing the effectiveness of the raters in overcoming random chance.

The philosophical foundation of the Kappa statistic insists that high agreement is only meaningful if it significantly exceeds chance. Imagine a scenario where 90% of raters agree that a product is "Poor." If, however, 90% of all items are rated "Poor" across the board due to inherent bias or poor product quality, then the 90% agreement might simply be an artifact of the category distribution, not true consensus on the specific items. Fleiss' Kappa corrects for this by penalizing agreements that are heavily driven by skewed marginal distributions. This adjustment ensures that a high Kappa value truly reflects a successful, repeatable, and non-random alignment of human judgment. This inherent robustness makes Fleiss' Kappa the preferred method for assessing reliability in studies involving multiple independent judges, such as clinical trials or complex scoring mechanisms where standardization of assessment protocols is paramount.

Data Preparation and Structure for Calculation

Before initiating the calculation of Fleiss' Kappa in Excel, meticulous data preparation is required. The data must be organized into a specific contingency table format. Each row of the table must correspond to a unique item being rated (N items), and the columns must correspond to the categories available for rating (K categories). The cell entries themselves must not contain the actual ratings (e.g., "Poor" or "Excellent") but rather the count of raters (out of the total number of raters, M) who assigned that specific item to that specific category. For instance, if 14 raters assess 10 products across 5 categories, the resulting matrix will have 10 rows and 5 data columns, plus a column summing the total ratings per item (which should always equal 14, the total number of raters). This matrix, often referred to as the $N times K$ matrix of counts, is the raw material for the analysis.

Suppose 14 individuals rated 10 different products using a scale from Poor to Excellent. This setup implies $N=10$ items and $M=14$ raters. If the scale consists of five levels (e.g., Poor, Fair, Average, Good, Excellent), then $K=5$. The data visualization provided illustrates exactly this structure, where the sum of counts across the categories for any given product is 14. This initial step of transforming raw ratings (which might initially be recorded as 10 rows of 14 ratings each) into the necessary count matrix is often the most time-consuming and critical part of the process. Errors in this conversion will propagate throughout the entire calculation, leading to an invalid Kappa statistic. It is essential to verify that the row totals equal M (the number of raters) before proceeding, as shown in the screenshot displaying the total ratings for each product:

	A	B	C	D	E	F	G	H	I
1			**Rating**						
2		**Product ID**	Poor	Weak	Average	Good	Excellent		
3		1	0	0	0	0	14		
4		2	0	2	6	4	2		
5		3	0	0	3	5	6		
6		4	0	3	9	2	0		
7		5	2	2	8	1	1		
8		6	7	7	0	0	0		
9		7	3	2	6	3	0		
10		8	2	5	3	2	2		
11		9	6	5	2	1	0		
12		10	0	2	2	3	7		
13									
14									
15									
16									
17									
18									
19									
20									
21									

Once the count matrix is correctly established, the next stage involves deriving the proportions necessary for the Fleiss' Kappa formula. Although standard Excel does not contain a dedicated built-in function for Fleiss' Kappa, the calculation can be achieved using a series of sequential column calculations involving standard arithmetic functions. The key is to calculate the necessary sums and squared sums. For example, for each item, we need to calculate the degree of agreement observed. This is done by looking at the counts in each category for that item, applying the necessary mathematical factor related to the number of raters, and summing these values. These intermediate results are then aggregated to find the overall observed agreement, P_o, and the expected agreement, P_e, which hinges on the marginal proportions of categories across the entire dataset. The successful execution of these steps relies entirely on the foundational count matrix being accurate.

Detailed Calculation Steps in Excel

The calculation of Fleiss' Kappa is best approached methodically, breaking down the complex formula into distinct, manageable steps executed in separate columns within the Excel sheet. This not only facilitates debugging but also makes the statistical logic transparent. The calculation process involves four main stages: 1) calculating the agreement component for each item, 2) finding the mean observed agreement (P_o), 3) calculating the marginal proportions for each

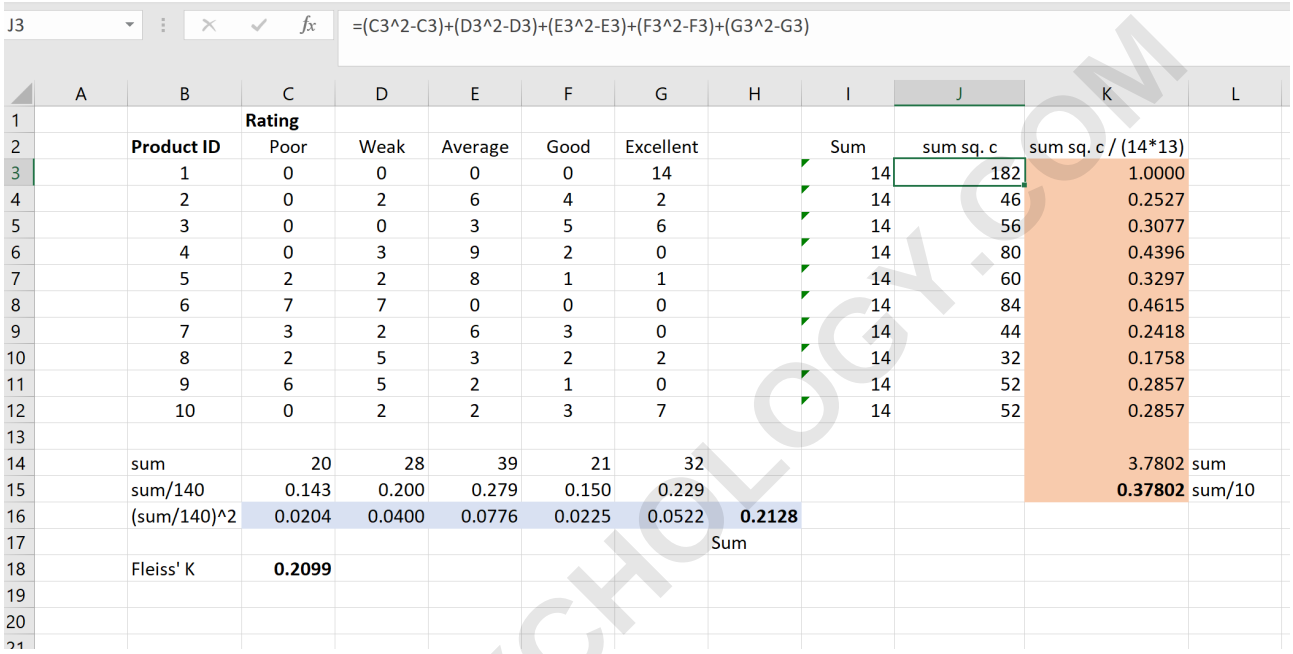
category, and 4) finding the mean expected agreement (P_e). The final Kappa value is then derived using these two aggregate measures. While the exact steps might vary slightly depending on the specific arrangement of the spreadsheet, the underlying mathematical principles remain constant, ensuring the reliability of the derived statistic.

The first crucial calculation step involves determining the degree of agreement for each item. For each item i, we calculate P_i, which is the proportion of all possible pairs of raters who agreed on that item. The numerator in this sub-calculation is the sum of products of counts for each category j, specifically $n_{ij}(n_{ij} - 1)$, summed across all categories K. This calculation effectively counts the number of agreeing pairs within the categories for that item. This sum is then divided by the total number of possible pairs of raters, $M(M-1)$, where M is the number of raters (14 in our example). This P_i value, calculated in a dedicated column, represents the raw observed agreement for item i, adjusted for the total number of possible pairs. The image provided suggests that this challenging calculation is consolidated, possibly in Column J, highlighting its complexity relative to simple sums or averages.

Following the calculation of the individual item agreements (P_i), the next step involves calculating the mean observed agreement, P_o. This is simply the arithmetic mean of all the P_i values calculated in the previous step, averaging the agreement across all N items. Mathematically, $P_o = (1/N) sum P_i$. This measure establishes the overall level of consensus achieved across the entire set of products. Simultaneously, we must calculate the expected agreement by chance, P_e. This requires first determining the marginal proportion for each category. For category j, the proportion p_j is calculated by summing the counts in category j across all N items and dividing by the total number of ratings ($N times M$). Once all marginal proportions ($p_1, p_2, ldots, p_K$) are determined, P_e is calculated by summing the squares of these marginal proportions: $P_e = sum p_j^2$. This squaring step is key, as it represents the probability of two random raters independently selecting the same category, based on the overall category usage rates.

The final stage is the assembly of the Fleiss' Kappa formula: $kappa = (P_o - P_e) / (1 - P_e)$. Excel users must ensure that they use parentheses correctly to enforce the order of operations, calculating the numerator and the denominator separately before performing the final division. Referring to the provided screenshot, the complexity of these calculations, especially the derivation of the P_i values (implied in column J), requires careful spreadsheet design. The final Kappa value, as demonstrated in the example, is obtained by substituting the calculated P_o and P_e values into the formula. For instance, the example provided shows an intermediate step of: Fleiss' Kappa = (0.37802 - 0.2128) / (1 - 0.2128) = 0.2099. This confirms that $P_o approx 0.37802$ (the mean observed agreement) and $P_e approx 0.2128$ (the mean expected agreement by chance). The resulting statistic, 0.2099, is the central output of the entire analysis, representing the measure of inter-rater reliability.

The visual representation below shows how these complex steps are typically mapped out in a functional Excel spreadsheet, culminating in the final calculation cell. Notice how intermediate calculations, crucial for deriving P_o and P_e, occupy the specialized columns, simplifying the final formula application in cell C18. The transparency offered by performing these calculations manually in Excel is highly valuable for didactic purposes, ensuring a deep understanding of the statistical mechanism at play.



J3 =(C3^2-C3)+(D3^2-D3)+(E3^2-E3)+(F3^2-F3)+(G3^2-G3)

	A	B	C	D	E	F	G	H	I	J	K	L
1			Rating									
2		Product ID	Poor	Weak	Average	Good	Excellent		Sum	sum sq. c	sum sq. c / (14*13)	
3		1	0	0	0	0	14		14	182	1.0000	
4		2	0	2	6	4	2		14	46	0.2527	
5		3	0	0	3	5	6		14	56	0.3077	
6		4	0	3	9	2	0		14	80	0.4396	
7		5	2	2	8	1	1		14	60	0.3297	
8		6	7	7	0	0	0		14	84	0.4615	
9		7	3	2	6	3	0		14	44	0.2418	
10		8	2	5	3	2	2		14	32	0.1758	
11		9	6	5	2	1	0		14	52	0.2857	
12		10	0	2	2	3	7		14	52	0.2857	
13												
14		sum	20	28	39	21	32				3.7802	sum
15		sum/140	0.143	0.200	0.279	0.150	0.229				**0.37802**	sum/10
16		(sum/140)^2	0.0204	0.0400	0.0776	0.0225	0.0522	**0.2128**				
17								Sum				
18		Fleiss' K	**0.2099**									
19												
20												

The Distinction: Fleiss' Kappa vs. Cohen's Kappa

A frequent point of confusion for researchers involves differentiating between Fleiss' Kappa and Cohen's Kappa. While both statistics serve the fundamental purpose of measuring agreement adjusted for chance, their applicability is strictly determined by the number of raters involved. Cohen's Kappa is designed exclusively for situations involving exactly two raters. It is a powerful tool for pairwise reliability assessment, particularly when the raters are observed over a series of independent judgments. However, when the study expands to include three, four, or more raters ($M > 2$), Cohen's Kappa becomes inadequate, as it cannot simultaneously model the consensus among all participants. This is precisely where Fleiss' Kappa steps in, providing the necessary generalization to accommodate any fixed number of raters, making it the standard choice for multi-rater reliability studies.

The calculation methodologies also differ subtly but significantly. Cohen's Kappa requires a contingency table that links the two raters' decisions item by item. It requires knowing who rated what. Fleiss' Kappa, however, does not require identifying which specific rater made which specific judgment. Instead, it works purely with the count of agreements per category for each item. This

distinction is crucial: Fleiss' Kappa assumes that the raters are **exchangeable**--meaning it doesn't matter which rater contributed which specific score, only how many raters chose a given category--while Cohen's Kappa treats the two raters as distinct entities whose individual agreement patterns are measured. This exchangeability assumption is essential for simplifying the calculation across multiple raters.

Because Fleiss' Kappa generalizes the concept of agreement across a group, it is often utilized in large-scale studies where many analysts independently evaluate a substantial body of data, such as content analysis or large medical reviews. While the underlying interpretation scale is often borrowed from Cohen's Kappa, it is important to remember that a moderate Fleiss' Kappa value (say, 0.50) reflects moderate agreement among an entire cohort (e.g., 14 raters), whereas a Cohen's Kappa of 0.50 reflects moderate agreement between just two individuals. Understanding this contextual difference is vital for accurately interpreting the reliability findings and drawing appropriate conclusions about the replicability of the rating process.

Interpreting the Fleiss' Kappa Statistic

Once the Fleiss' Kappa value is successfully calculated, the final and most important step is interpreting the result within the context of the study. Unlike many other statistics, there is no single universally accepted, formal set of rigid criteria for interpreting the magnitude of Kappa. However, statistical researchers often rely on established guidelines, such as those proposed by Landis and Koch (1977) or Altman (1991), typically adapted from the scales used for Cohen's Kappa. These guidelines provide a standardized framework for assessing the practical significance of the calculated reliability score, allowing researchers to categorize the level of agreement achieved.

The standard reference scale often used to interpret Kappa values is structured as follows, providing clear cutoffs for reliability levels. Although this scale was originally developed for Cohen's Kappa, it is widely used as a heuristic guide for Fleiss' Kappa interpretation, acknowledging that it gives a useful approximation of the strength of agreement:

< 0.20 | Poor agreement
0.21 - 0.40 | Fair agreement
0.41 - 0.60 | Moderate agreement
0.61 - 0.80 | Good agreement
0.81 - 1.00 | Very Good agreement

Applying this interpretation scale to the result from our Excel example, where the calculated Fleiss' Kappa was determined to be **0.2099**, we can classify the outcome. Since 0.2099 falls within the $0.21 - 0.40$ range, albeit marginally, the level of inter-rater agreement among the 14 product raters is considered "Fair." This finding suggests that while the agreement observed is better than random chance, there is still significant variability or lack of consensus among the raters. A "Fair"

rating often indicates that the rating protocol or the training of the raters may need refinement, or that the categorical ratings themselves may be insufficiently distinct, leading to inconsistent subjective judgments.

It is important for content creators and editors to clearly state that while these scales are useful, interpretation must always be contextual. A "Good" Kappa (0.70) might be considered excellent in clinical diagnostics where subjective assessment is inherently difficult, but perhaps only acceptable in highly controlled mechanical settings. Conversely, a "Poor" rating (e.g., 0.15) should prompt a mandatory reassessment of the study methodology, the rater training, or the category definitions themselves, as the data collected under such low reliability cannot be confidently generalized. The goal is always to maximize the Kappa score, aiming for values in the "Good" or "Very Good" range to ensure the highest possible validity and reliability of the research findings derived from the collected data.

Conclusion: Achieving Reliable Ratings

Calculating Fleiss' Kappa in Excel, while demanding rigorous attention to detail and precise formula application, provides researchers with a robust method for validating the consistency of multi-rater judgments. The process moves beyond simple agreement percentages by rigorously controlling for the role of chance, yielding a statistic that truly reflects the systematic reliability inherent in the rating procedure. Whether dealing with 3 raters or 30, the ability to calculate P_o (observed agreement) and P_e (expected agreement) and combine them into the final Kappa formula is a cornerstone of sound quantitative analysis.

The example demonstrated a final Fleiss' Kappa of **0.2099**, which signifies a "fair" level of agreement among the 14 raters. This result is instructive: it confirms that while Excel can be utilized effectively for these calculations, the resulting reliability score itself often serves as a critical feedback mechanism. A "fair" score necessitates further investigation into the rating process. Researchers should ask: Were the criteria clear enough? Was the training consistent? Could the scale categories be improved? Only by addressing these methodological questions can a research team strive toward achieving the "Good" or "Very Good" reliability levels that underpin high-quality, trustworthy data acquisition.

In summary, mastering the steps for calculating Fleiss' Kappa in Excel ensures that researchers can effectively diagnose the reliability of their data collection methods involving human judgment. This statistical rigor is not merely an academic exercise; it is a practical necessity for ensuring that conclusions drawn from subjective data--whether in product evaluation, psychological assessment, or medical diagnosis--are reliable, reproducible, and robust against the influence of random error.