

# What is Inter-rater Reliability?

Authored by  
**stats writer**

April 26, 2024

## RECOMMENDED CITATION

stats writer (2024). *What is Inter-rater Reliability?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=139672>

Inter-rater reliability refers to the consistency or agreement between multiple raters or observers in their evaluations or measurements of a particular phenomenon. It is a statistical measure that assesses the level of agreement between different individuals who are tasked with observing or assessing the same thing. Inter-rater reliability is important in research, as it ensures that the results obtained are not influenced by individual bias or subjectivity. For example, in a study evaluating the effectiveness of a new teaching method, inter-rater reliability would be assessed by having multiple teachers independently observe and rate students' performance using the new method. This would help to ensure that the results are not influenced by the personal opinions or biases of a single teacher.

## **What is Inter-rater Reliability? (Definition & Example)**

**In statistics, inter-rater reliability is a way to measure the level of agreement between multiple raters or judges.**

**It is used as a way to assess the reliability of answers produced by different items on a test. If a test has lower inter-rater reliability, this could be an indication that the items on the test are confusing, unclear, or even unnecessary.**

**There are two common ways to measure inter-rater reliability:**

### **1. Percent Agreement**

**The simple way to measure inter-rater reliability is to calculate the percentage of items that the judges agree**

on.

This is known as percent agreement, which always ranges between 0 and 1 with 0 indicating no agreement between raters and 1 indicating perfect agreement between raters.

For example, suppose two judges are asked to rate the difficulty of 10 items on a test from a scale of 1 to 3. The results are shown below:

	Judge 1	Judge 2
Question 1	1	1
Question 2	1	1
Question 3	2	3
Question 4	2	2
Question 5	1	2
Question 6	2	3
Question 7	3	3
Question 8	2	2
Question 9	3	3
Question 10	3	3

For each question, we can write "1" if the two judges agree and "0" if they don't agree.

	Judge 1	Judge 2	Agree?
Question 1	1	1	1
Question 2	1	1	1
Question 3	2	3	0
Question 4	2	2	1
Question 5	1	2	0
Question 6	2	3	0
Question 7	3	3	1
Question 8	2	2	1
Question 9	3	3	1
Question 10	3	3	1

The percentage of questions the judges agreed on was  $7/10 = 70\%$ .

## 2. Cohen's Kappa

The more difficult (and more rigorous) way to measure inter-rater reliability is to use Cohen's Kappa, which calculates the percentage of items that the raters agree on, while accounting for the fact that the raters may happen to agree on some items purely by chance.

The formula for Cohen's kappa is calculated as:

$$k = (p_o - p_e) / (1 - p_e)$$

where:

**po: Relative observed agreement among raters**

**pe: Hypothetical probability of chance agreement**

**For a step-by-step example of how to calculate Cohen's Kappa, refer to .**

#### **How to Interpret Inter-Rater Reliability**

**The higher the inter-rater reliability, the more consistently multiple judges rate items or questions on a test with similar scores.**

**In general, an inter-rater agreement of at least 75% is required in most fields for a test to be considered reliable. However, higher inter-rater reliabilities may be needed in specific fields.**

**For example, an inter-rater reliability of 75% may be acceptable for a test that seeks to determine how well a TV show will be received.**

**On the other hand, an inter-rater reliability of 95% may be required in medical settings in which multiple doctors are judging whether or not a certain treatment should be used on a given patient.**

**Note that in most academic settings and rigorous fields of research, Cohen's Kappa is used to calculate inter-rater reliability.**

ARABPSYCHOLOGY.COM