# ICC vs Kappa vs Bland Altman

Answer two visible questions about your data below to find the one matching measure. Nothing is inferred from data you have not entered; every route is based only on your two answers.

**StatReason**

ICC vs Kappa vs Bland Altman



**1. What kind of data are you comparing?**

**2. How many raters or methods?**

## Recommended measure: Bland-Altman Calculator

Go to the Bland-Altman Calculator →

## What this answers

This tool answers "which specific reliability or agreement measure fits my two-question situation?" It uses only the type of data you are comparing and how many raters or methods are involved, since those two facts determine the one correct measure among ICC, kappa, weighted kappa, and Bland-Altman.

## Why data type and rater count decide the measure

Continuous numeric measurements from two methods that should agree closely call for Bland-Altman analysis, which shows the average difference and the limits of agreement between two

measurement approaches. Continuous measurements from three or more raters call for the intraclass correlation coefficient, which quantifies how much of the total variance is due to genuine differences between subjects versus rater inconsistency. Categorical ratings from two raters call for Cohen's kappa; ordered categorical ratings, where a near-miss should count differently from a complete disagreement, call for weighted kappa instead of the unweighted version.

## Worked example

Comparing a new blood pressure device against a hospital-standard device on the same 40 patients is continuous data from two methods, routing to Bland-Altman analysis to check whether the two devices agree closely enough for interchangeable use. If instead three independent nurses each measured the same 40 patients with the same device, that would be continuous data from three raters, routing to the intraclass correlation coefficient instead, since the question shifts from "do two methods agree" to "how consistent are multiple raters using the same method."

#### Assumption audit

> **Calculated from your data:** nothing here; this page only routes based on the two answers you select and performs no calculation itself.

> **Evidence to review:** once you reach the recommended engine, its own Assumption Audit checks the specific facts, such as scale ordering or rater count, relevant to that measure.

> **You must verify:** whether your categorical scale is genuinely ordered before choosing weighted kappa over the unweighted version, since that distinction changes how disagreements are penalized.

## Source

This routing structure follows the reliability and agreement measure selection guidance in the NIST/SEMATECH e-Handbook of Statistical Methods and the shared statistical reasoning contract every StatReason engine is built against.

## Limitations

This tool covers the most common two-question routing case; a mixed design with both categorical and continuous elements, or split-half reliability for test items, is covered separately on its own engine page.