Reliability vs Agreement

Reliability, agreement, and validity are three related but distinct questions about a measurement, and only one of them is what most reliability statistics actually measure.

What this answers

This page answers "is Cronbach's alpha the same thing as inter-rater agreement, and does a high score on either one mean my measurement is valid?" Reliability, agreement, and validity each answer a separate question, and confusing them leads to overstated claims about what a measurement actually accomplishes.

Reliability: internal consistency

Reliability measures such as Cronbach's alpha or split-half reliability ask whether multiple items intended to measure the same underlying construct produce consistent results with each other. A high reliability score means the items hang together statistically; it says nothing about whether they are measuring the construct you actually intended to measure.

Agreement: do raters or methods concur

Agreement measures such as Cohen's kappa, weighted kappa, or the intraclass correlation coefficient ask a different question: do two or more raters, or two measurement methods, produce the same or similar results for the same subjects. Bland-Altman analysis specifically compares two measurement methods to see whether they agree closely enough to be used interchangeably, and by how much they might systematically differ.

Neither implies validity

Validity is the separate question of whether a measurement actually captures the concept it claims to measure. A scale can be highly reliable, with items that consistently agree with each other, while still measuring the wrong construct entirely, or two raters can agree closely on a flawed rating scheme. High reliability or agreement is often necessary for a measurement to be useful, but it is never sufficient on its own to establish validity.

Worked example

A 10-item customer satisfaction survey has a Cronbach's alpha of 0.88, indicating the items are internally consistent with each other. Two independent raters scoring open-ended responses on a 5-point scale have a weighted kappa of 0.72, indicating substantial but not perfect agreement. These are two separate findings about two separate aspects of the measurement process, and neither one confirms that the survey or the rating scale actually measures customer satisfaction accurately.

Assumption audit

> **Calculated from your data:** the specific reliability or agreement coefficient, once you run the relevant linked calculator on your item scores or rater ratings.

> **Evidence to review:** whether your data structure matches the measure's intended use, continuous items for Cronbach's alpha, categorical ratings for kappa, or paired continuous measurements for Bland-Altman.

> **You must verify:** the validity of your measurement instrument separately, through content review or comparison against an external standard, since no reliability or agreement statistic can establish it.

Source

This distinction follows the standard reliability and agreement framework in the NIST/SEMATECH e-Handbook of Statistical Methods and the shared statistical reasoning contract every StatReason engine is built against.

Limitations

This page introduces the conceptual distinction; choosing the specific correct measure for your data type and rater count is covered in ICC vs Kappa vs Bland Altman.

Next action: check internal consistency with **Cronbach's Alpha Calculator**, or rater agreement with **Cohen's Kappa Calculator**.