How to Easily Understand and Calculate the Intraclass Correlation Coefficient (ICC)

Authored by

stats writer

December 6, 2025

The Core Concept of the Intraclass Correlation Coefficient (ICC)

The **Intraclass Correlation Coefficient** (ICC) is a highly specialized statistical measure used extensively to quantify the degree of similarity or **reliability** among measurements or ratings provided by multiple observers, often referred to as **raters** or judges. This coefficient is essential in studies where two or more individuals independently evaluate the same set of subjects or items using a continuous or ordinal scale. Fundamentally rooted in the Analysis of Variance (ANOVA) framework, the ICC effectively partitions the total variance in the observed scores into components attributable to differences between the subjects themselves and differences introduced by the measurement process or the raters.

The application of the ICC is critical for ensuring that research findings possess strong measurement integrity. By quantifying inter-rater reliability, researchers can determine whether variations in scores reflect genuine differences among the items being measured or merely systematic or random error introduced by the observers. A high ICC value is the goal, indicating that the measurement tool and the **raters** are dependable, thus maximizing confidence in the data collected. The primary function of the ICC is to answer the crucial research question: Can different observers apply a specific measurement protocol consistently to yield comparable results?

The Range and Fundamental Interpretation of the ICC

The calculated value of the ICC is mathematically constrained to range from 0 to 1, offering a straightforward index of measurement quality. This range provides a standardized scale for interpreting the level of agreement observed within the study population. Specifically, an ICC value approaching 0 indicates poor **reliability**, suggesting that the variation between the ratings is substantial, potentially exceeding the true variation between the subjects being rated. In such a case, the measurement process is highly inconsistent.

Conversely, an ICC value near 1 signifies exceptional or perfect reliability. This high score confirms that the differences in scores observed are almost entirely attributable to the inherent differences among the subjects, demonstrating strong agreement and consistency among the multiple observers. Researchers must use established guidelines, detailed later, to determine what constitutes an acceptable level of reliability for their specific domain, as relying solely on the extreme points of the range (0 and 1) is rarely sufficient for practical decision-making.

Three Critical Factors Defining the ICC Calculation

The calculation of the **Intraclass Correlation Coefficient** is complex because it must be precisely tailored to the experimental design of the reliability study. There are several different versions of the ICC that can be computed, and the specific version used depends upon three critical

methodological decisions made by the researcher. These decisions reflect assumptions about the sampling methods used for both subjects and **raters**, as well as the intended use of the resulting reliability measure. Correctly specifying these factors is necessary to ensure the derived ICC is a valid reflection of the desired measure of agreement.

These defining choices govern how variance components--such as subject variance, rater variance, and residual error--are incorporated into the numerator (representing agreement) and the denominator (representing total variance) of the ICC formula. The three determinative factors are:

Model: This specifies the relationship between subjects and raters, selecting between the One-Way Random Effects, Two-Way Random Effects, or Two-Way Mixed Effects approaches.
Type of Relationship: This defines what constitutes agreement, distinguishing between **Consistency** (relative ranking agreement) or **Absolute Agreement** (score proximity).
Unit: This clarifies the intended use of the measurement in future applications, focusing on the reliability of a **Single rater** or the **Mean of raters**.

Understanding ICC Models: Fixed vs. Random Effects

The choice of ICC model is dictated by the sampling procedure used for the **raters** and determines whether the findings can be generalized to a broader population of observers. The models differentiate between fixed effects (where the raters used are the only ones of interest) and random effects (where the raters are a random sample meant to represent a larger population).

The **One-way random effects model** is applicable when each subject in the study is assessed by a unique and randomly chosen group of raters. In this scenario, the variation among raters is pooled into the error term, as the specific identity of the rater changes for every subject. Because practical inter-rater studies typically employ the same set of observers to rate all subjects, this model is rarely selected in applied research settings.

The **Two-way random effects model** is the most frequently employed model when assessing generalizability. It assumes that a fixed number of *k* raters were randomly selected from a larger population of potential raters, and this same group evaluated every subject. Both the subjects and the raters are treated as sources of random effects. This model is ideal when the research goal is to extrapolate the observed **reliability** results to any other set of similarly qualified raters who might utilize the measurement tool in the future.

Conversely, the **Two-way mixed effects model** is selected when the specific group of *k* raters chosen for the study is considered a fixed set. This model posits that these particular raters are the only ones relevant to the measurement process, and the researcher has no interest in generalizing the reliability findings to any other raters. This structure is typically used when assessing the consistency of a highly specialized and fixed team, such as a panel of expert clinicians or

researchers.

Distinguishing Between Consistency and Absolute Agreement

The 'Type of Relationship' factor defines the stringentness of the agreement required between the raters' scores. This methodological choice directly impacts how differences between observers are statistically treated when calculating the ICC.

Consistency: This measure focuses on the systematic relationship between the ratings provided by judges. It assesses whether raters maintain the same relative ranking of the subjects, even if one rater consistently assigns higher scores than another. For instance, if Judge A always rates subjects 1 point higher than Judge B, the correlation for consistency would be high because their ratings are systematically consistent in their relative placement of subjects. Consistency is concerned with proportional and ordinal agreement.

Absolute Agreement: This is a significantly more conservative measure of reliability, demanding that the raw scores provided by different judges for the same subject be numerically close or identical. Absolute differences, including any systematic bias (e.g., one judge being consistently harsher than another), are treated as error variance. This measure is essential when the precise magnitude of the score is clinically or statistically meaningful, and systematic discrepancies cannot be tolerated.

Choosing the Appropriate Unit of Measurement (Single vs. Mean)

The final factor, 'Unit,' addresses the practical question of how the resulting reliability measure will be applied in subsequent research or clinical practice. The choice is based on whether future scores will be derived from a single observation or averaged across multiple observations.

Single Rater: If the eventual measurement of the subjects will rely on the rating provided by just one, arbitrarily selected **rater**, then the single unit is appropriate. The resulting ICC reflects the expected reliability of any individual measurement taken by a single observer chosen from the population of raters.

Mean of Raters: When the research protocol dictates that the definitive score for each subject must be the average (mean) of the ratings provided by all *k* judges, the mean unit is selected. Calculating the mean inherently smooths out random individual errors, meaning the ICC calculated using the mean of raters will almost always yield a higher estimate of **reliability** compared to the single rater unit for the same dataset.

Note: *If you are interested in measuring the level of agreement between only two raters who classify items using nominal or categorical scales, you should instead use the Cohen's Kappa*

coefficient, as the ICC is designed specifically for continuous or ordinal data rated by two or more observers.

Practical Guidelines for Interpreting ICC Scores

Once the **Intraclass Correlation Coefficient** is calculated, its value must be interpreted using accepted academic standards to determine the suitability of the measurement instrument. While interpretation can be discipline-specific, the following widely accepted benchmarks, or rules of thumb, provide a consistent framework for assessing the quality of the observed reliability:

Less than 0.50: Indicates **Poor reliability**, suggesting the measurement is unreliable and likely invalid.
Between 0.5 and 0.75: Suggests **Moderate reliability**, which may be acceptable for pilot studies or exploratory research, but not ideal for definitive clinical measurements.
Between 0.75 and 0.9: Demonstrates **Good reliability**, sufficient for most established research applications and many clinical settings.
Greater than 0.9: Represents **Excellent reliability**, indicating a high degree of confidence in the measurement process and minimal influence from rater variability.

Calculating the ICC: A Practical Example in R

Consider a scenario where the quality of 10 different college entrance exams (subjects) is rated by four independent judges (**raters**). The scores, presumably on a continuous scale, are recorded in the data table below, allowing us to compute the ICC based on the design parameters chosen.

Exam	Judge A	Judge B	Judge C	Judge D
1	1	2	0	1
2	1	3	4	2
3	3	8	1	3
4	6	4	5	3
5	6	5	5	6
6	7	5	6	4
7	8	7	6	6
8	9	9	9	8
9	8	8	8	8
10	7	8	8	9

For our calculation, we assume the following critical choices: The four judges were randomly

sampled from a larger population of qualified examiners (suggesting generalization is desired). We are interested in measuring the **absolute agreement** among them, meaning we want their scores to be numerically close. Finally, we anticipate that in future applications, a single judge's assessment will be used as the final score, so we require the **single rater** unit perspective.

Based on these assumptions, we must execute a two-way random effects model using the absolute agreement type and the single rater unit. The following code demonstrates the calculation of the ICC using the `irr` package in the statistical software R:

```
#load the interrater reliability package
library(irr)

#define data
data <- data.frame(A=c(1, 1, 3, 6, 6, 7, 8, 9, 8, 7),
B=c(2, 3, 8, 4, 5, 5, 7, 9, 8, 8),
C=c(0, 4, 1, 5, 5, 6, 6, 9, 8, 8),
D=c(1, 2, 3, 3, 6, 4, 6, 8, 8, 9))

#calculate ICC
icc(data, model = "twoway", type = "agreement", unit = "single")

Model: twoway
Type : agreement

Subjects = 10
Raters = 4
ICC(A,1) = 0.782

F-Test, H0: r0 = 0 ; H1: r0 > 0
F(9,30) = 15.3 , p = 5.93e-09

95%-Confidence Interval for ICC Population Values:
0.554 < ICC < 0.931
```


The output of the statistical analysis reveals that the **Intraclass Correlation Coefficient** (ICC) is calculated to be **0.782**. This value is derived using the variance components that specifically account for a two-way random design demanding absolute agreement based on a single measurement. The accompanying F-test further validates this result by showing that the observed reliability is highly significant, confirming it is statistically greater than zero.

Interpreting the calculated ICC of **0.782** against the established guidelines, we find that it falls within the range categorized as "Good reliability." Therefore, we can confidently conclude that the

quality rating process for these exams is robust and dependable. The measurement instrument allows different, randomly selected **raters** to achieve a high degree of absolute agreement when assessing the exams, making the resulting scores suitable for research or high-stakes evaluation purposes.

Further Resources on Intraclass Correlation

For researchers seeking deeper technical understanding regarding the mathematical underpinnings of the ICC, or detailed instructions on its computation in other statistical platforms like SPSS or Stata, the following tutorials and documentation resources are recommended for in-depth study.