Learning Guide: Calculating the Intraclass Correlation Coefficient (ICC) in R

Authored by

Mohammed looti

November 5, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Learning Guide: Calculating the Intraclass Correlation Coefficient (ICC) in R*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=10599

The Intraclass Correlation Coefficient (ICC) stands as a fundamental statistical measure utilized primarily to quantify the degree of resemblance or **reliability** among multiple measurements or ratings applied to the same set of subjects. In fields ranging from medical research to educational psychology, assessing whether judges, observers, or measurement instruments can consistently rate items is essential, making the ICC a critical tool for establishing the trustworthiness of data collection processes.

Unlike traditional Pearson correlation coefficients, which gauge the linear relationship between two distinct variables, the ICC is specifically engineered for clustered data structures--situations where multiple raters evaluate the same subjects. This coefficient provides a standardized index indicating what proportion of the total observed variance in the scores is genuinely attributable to differences between the subjects being rated, rather than being mere noise or variation caused by the raters themselves.

The resulting ICC value is constrained between 0 and 1. A score approaching 0 signifies negligible similarity or **reliability**, implying that the variance observed is almost entirely due to rater discrepancies or random error. Conversely, an ICC value near 1 denotes perfect **reliability**, meaning that all raters assign virtually identical scores to the same subjects. A robust ICC is a prerequisite for validating assessment tools and ensuring that research findings are generalizable and dependable.

Leveraging the R Environment and the irr Package for ICC Analysis

To calculate the Intraclass Correlation Coefficient efficiently and accurately, the standard approach within the R programming language utilizes specialized libraries. The most widely adopted resource for this task is the **irr** package, an acronym for Interrater Reliability. This robust package significantly simplifies the complex variance component calculations required for various ICC models, providing a streamlined workflow for researchers.

The core function employed for this analysis is **icc()**. Proper execution of this function demands specific input parameters that must align precisely with the underlying study design and the statistical assumptions being made about the data. Understanding and correctly specifying these arguments is paramount to deriving a meaningful and valid reliability estimate.

The general syntax for invoking the **icc()** function follows this structure:

icc(ratings, model, type, unit)

A detailed comprehension of each required argument is essential for successful implementation:

ratings: This foundational input must be a data frame or matrix containing the observed scores. Conventionally, each column represents the scores assigned by a specific rater, while each row

corresponds to a single subject or item that was rated.
model: This crucial parameter specifies the statistical assumptions concerning the selection of raters and subjects. Options include the **"oneway"** (One-Way Random Effects) or the **"twoway"** (Two-Way Random or Mixed Effects) model, depending on whether raters are fixed or randomly sampled.
type: Defines the specific concept of reliability being measured. The **"consistency"** type assesses whether raters rank subjects similarly, whereas the **"agreement"** type measures the degree of absolute agreement, requiring the actual numerical scores to be nearly identical.
unit: Determines the scope of the resulting reliability estimate. Setting it to **"single"** calculates the reliability of one typical rater, while **"average"** estimates the reliability if the mean score across all raters were used as the measure.

Step 1: Structuring and Preparing Data for ICC Calculation

The first necessary step before executing the coefficient calculation is ensuring the data is correctly structured within the R programming language environment. The **irr** package requires input data to be formatted as a matrix where the rows represent the items or subjects being assessed, and the columns represent the individual raters or measurement attempts.

To illustrate this process, let us consider a practical example: evaluating the reliability of scoring ten distinct college entrance exams. Four independent judges (designated Raters A, B, C, and D) were commissioned to assign a numerical quality rating to each exam. Our dataset must reflect these ten subjects and four raters accurately.

This specific matrix organization is crucial because the ICC calculation relies on analyzing the variance components across both the rows (subjects) and the columns (raters). We construct this required data frame in R using the following script, where the scores provided reflect the individual assessments by each judge:

```
#create 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))
```

Step 2: Defining the Statistical Model and Reliability Type

Selecting the correct statistical model is the most pivotal decision in calculating the ICC, as it directly reflects the design of the study and the scope of inference desired. In our scenario, since the four judges were randomly sampled from a larger population of qualified assessors and the goal is to generalize the findings back to that population, the **Two-Way Random Effects Model** is the methodologically appropriate choice.

Next, we must specify the type of relationship we are measuring. If the research question focuses solely on whether raters maintain similar rankings (consistency), one parameter would be used. However, if the study requires that raters assign scores that are numerically close--that is, demanding absolute agreement--the parameters must be adjusted accordingly. For this example, we prioritize measuring **absolute agreement**, ensuring the magnitude of the scores, not just the relative ranking, is consistent across judges.

Finally, the unit parameter must be set. Because our objective is to estimate the reliability inherent in a score derived from any single, randomly selected judge, we specify the unit parameter as **"single."** These three interconnected choices--Model ("twoway"), Type ("agreement"), and Unit ("single")--collectively determine the exact ICC formula calculated, often recognized in literature as ICC(A,1).

Step 3: Executing the ICC Calculation and Reviewing the Output

With the data prepared and the necessary parameters established, the calculation phase begins. It is essential to first load the necessary irr package into the R session using the **library()** command. This makes the specialized ICC functions available for use.

The **icc()** function is then applied to the data frame. Based on our methodological decisions from Step 2, we must explicitly set the arguments: **model = "twoway"**, **type = "agreement"**, and **unit = "single"**. This specific configuration prompts R to calculate the reliability estimate based on the Two-Way Random Effects Model for Absolute Agreement, referencing a single measurement.

The following script consolidates the necessary commands--loading the package, defining the sample data for context, and executing the final ICC calculation--resulting in the comprehensive statistical output shown below:

```
#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

Interpreting the Intraclass Correlation Coefficient Result

Upon execution, the calculation provides the point estimate for the Intraclass Correlation Coefficient (ICC), which is determined to be **0.782**. This value is a crucial metric, quantifying the proportion of the total variability observed in the scores that is genuinely attributable to true differences between the rated subjects (the exams), while discounting error variance stemming from measurement inconsistencies or rater disagreement.

The output also includes a powerful test of statistical significance--the F-Test. The null hypothesis (H0: r0 = 0) posits that the true population ICC is zero, implying a complete lack of reliability. Given the highly favorable F-statistic of 15.3 and the extremely low p-value (p = 5.93e-09), we have compelling statistical evidence to reject the null hypothesis. This rejection confirms that the calculated reliability is highly significant and reliable, meaning it is unlikely to have occurred simply by chance.

Furthermore, the output furnishes a 95% Confidence Interval for the population ICC, spanning from 0.554 to 0.931. This interval suggests that if the study were replicated numerous times, the true reliability value in the broader population would fall within this range 95% of the time. While this range indicates inherent sampling uncertainty, it strongly supports the conclusion that the true inter-rater reliability is substantial, falling within the moderate to excellent range.

Contextualizing Reliability: Standard Interpretation Guidelines

To properly judge the strength of the calculated ICC value (0.782), researchers commonly refer to established benchmarks for interpreting reliability coefficients. Although these guidelines can sometimes vary across disciplines, they offer a standard, useful framework for categorizing the quality of inter-rater agreement based on the magnitude of the ICC score:

Less than 0.50: Classified as **Poor reliability**. Such a low score indicates critical issues with consistency or agreement, suggesting the assessment tool or raters are unreliable.
Between 0.50 and 0.75: Indicates **Moderate reliability**. This level is often deemed acceptable for exploratory or preliminary research, but improvements are usually recommended for high-stakes or definitive assessment instruments.
Between 0.75 and 0.90: Represents **Good reliability**. As demonstrated by our result of 0.782, this range is typically considered sufficient and robust for general research and many clinical applications.
Greater than 0.90: Signifies **Excellent reliability**. This score indicates near-perfect consistency and agreement, providing the highest level of confidence in the measurement process.

Based on this framework, our analysis confirms that the four independent judges exhibited a good level of absolute agreement when assessing the college entrance exams.

Methodological Depth: Advanced Considerations for ICC Selection

It is vital for analysts to understand that the term "Intraclass Correlation Coefficient" functions as an umbrella category encompassing numerous specific statistical formulas. The selection of the mathematically appropriate formula is not a matter of preference but a fundamental methodological requirement dictated by the specific research question and the exact experimental design used during data collection. Choosing the wrong model can lead to severely biased or misinterpreted reliability estimates.

To correctly pinpoint the specific ICC calculation required, three critical methodological factors must be determined based on the study design:

Model Definition: Are the raters considered fixed (e.g., specific, pre-selected experts, requiring a Two-Way Mixed Effects Model) or are they considered randomly sampled from a larger population (necessitating a One-Way or Two-Way Random Effects Model)?
Targeted Relationship: Does the study seek to measure how similarly raters rank subjects (Consistency), or does it require them to achieve the same numerical magnitude of scores (Absolute Agreement)?
Unit of Reliability: Is the final reliability estimate intended to apply to the score provided by a **Single** rater, or to the more stable reliability achieved by averaging the scores of all raters (the **Mean**)?

In the detailed example provided throughout this article, the analysis was rigorously based on the following methodological parameters, which resulted in the specific ICC(A,1) formula:

Model: Two-Way Random Effects
Type of Relationship: Absolute Agreement

Unit: Single rater

For researchers requiring a comprehensive theoretical exploration of the distinctions between the One-Way, Two-Way Mixed, and Two-Way Random Effects models, particularly how these differences impact the interpretation of consistency versus agreement, further consultation of specialized statistical texts is highly recommended.