Journal of Biostatistics and Epidemiology

*J Biostat Epidemiol. 2021;7(3): 305-309*

Original Article

# Guideline for Selecting Types of Reliability and Suitable Intra-Class Correlation Coefficients in Clinical Research

Khalil Taherzadeh Chenani, MSc[1], Farzan Madadizadeh, PhD[2*]

[1]Department of Occupational Health Engineering, School of Public Health, Occupational Health Research Center, Yazd, Iran.
[2]Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

ARTICLE INFO

Received 15.05.2021
Revised 28.05.2021
Accepted 24.06.2021
Published 25.09.2021

**Key words:**
Reliability;
Inter-rater reliability;
Test-retest reliability;
Intra-rater reliability;
Intra-cluster correlation coefficient

ABSTRACT

**Introduction:** Reliability is an integral part of measuring the reproducibility of research information. Intra-cluster correlation coefficient (ICC) is one of the necessary indicators for reliability reporting, which can be misleading in terms of its diversity. The main purpose of this study was to introduce the types of reliability and appropriate ICC indices.
**Methods:** In this tutorial article, useful information about the types of reliability and indicators needed to report the results, as well as the types of ICC and its applications were explained for dummies.
**Results:** Three general types of reliability include inter-rater reliability, test-retest reliability, and intra-rater reliability was presented. 10 different types of ICC were also introduced and explained.
**Conclusion:** The research results may be misleading if any of the reliability types and calculation criteria types are chosen incorrectly. Therefore, to make the results of the study more accurate and valuable. Medical researchers must seek help from relevant guidelines such as this study before conducting reliability analysis.

## Introduction

Reliability has remained one of the main concerns of clinical research in terms of consistency and potential sources of error in data analysis.[1] Reliability is the "stability of measurement over a variety of conditions in which basically the same results should be obtained".[2] It is also applied in different forms such as scale/tools reliability, rater/observer reliability, and response reliability.[3] Generally, there are three different types of reliability which include inter-rater reliability, test-retest reliability, and intra-rater reliability. This study aims to provide a short guideline for choosing different types of reliability and suitable types of intra-class correlation (ICC) in medical research. Therefore, the necessary parts are

*.Corresponding Author: madadizadehfarzan@gmail.com

Copyright © 2021 Tehran University of Medical Sciences. Published by Tehran University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

presented in the following order.

## 1-Reliability

### *1-1. Inter-rater reliability*

Inter-rater reliability (IRR), also called the inter-rater agreement, inter-rater concordance, and inter-observer reliability is defined as something that "reflects the variation between 2 or more raters who measure the same group of subjects".[4] In other words, this type of reliability assessment measures the consensus among the ratings given by different raters or observers. Joint-probability of agreement, Cohen's Kappa, Scott's pi, Fleiss' Kappa (Fleiss' K), Light's Kappa (for nominal variables), concordance correlation coefficient (for continuous variable), ICC (for continuous variable), polychoric correlation (for continuous variable), Gwet's AC1 and AC2 (dependent or independent raters), and Krippendorff's alpha (for all type of variables) are some statistical measures that can be applied for measuring IRR.[5] Additionally, Kendall correlation coefficient, Cohen's Weighted Kappa, Wilcoxon signed ranks test and sign test can be used for analyzing ordinal variables.[6] All indicators can be easily calculated in statistical software, for example, the "IRR" and "rel" package in R software can calculate IRR measures.

### *1-2. Test-retest reliability*

Test-retest reliability is defined as something that "reflects the variation in measurements taken by an instrument on the same subject under the same conditions. It is generally indicative of reliability in situations when raters are not involved or rater effect is negligible, such as self-report survey instrument".[4] In other words, this type of reliability assessment measures the closeness of the agreement between the results of trials that are taken by a single person or instrument in the same conditions.[7] ICC is indicated as the preferable index for measuring test-retest reliability.[8]

### *1-3. Intra-rater reliability*

Intra-rater reliability is defined as something that "reflects the variation of data measured by 1 rater across 2 or more trials".[4] In other words, this type of reliability assessment measures the degree of consistency among ratings given by one individual across multiple trials. Cohen's kappa (for nominal variable)[9], Fleiss' kappa (nominal/ordinal variable)[10] and ICC (for continuous variable)[11] are some indices that could be used for calculating the Intra-rater reliability.

## 2- Type of ICC

ICC is a descriptive index in statistics and is used when quantitative measurements are made based on units in classes, so it measures the similarity of the same class units. Its other application is for measuring the stability and reliability of quantitative scales. [5]

Unfortunately, many researchers do not state the type of ICC used in their research, and this is a methodological flaw because ICC has different types. At least 10 kinds of ICC are found in the literature.[5] Researchers should choose the appropriate type of ICC for their study because each of these involves specific assumptions and leads to different interpretations of the gathered data.[4]

McGraw and Wong have defined 10 kinds of ICC based on their "Model" (One-way random effects, Two-way random effects, or Two-way Mixed-effects), "Number of rate/ measurement" (single rater/ measurement or the mean of k-raters/measurements), and "type" (consistency or absolute agreement)[12] (See Figure 1).

### *Model*

- One-Way Random-Effects: In this model, different raters, randomly chosen from a larger population, rate the considered items.[4]
- Two-Way Random-Effects: This model is used when raters are randomly selected from a larger population of raters with similar characteristics.[4] In other words, Two-Way Random-Effects are selected when generalizing the reliability of the results is supposed.[4]
- Two-Way Mixed-Effects: This model is used when the raters are the raters of interest.[4] Which means that the reliability of the results cannot be generalized to other raters even with the same characteristics.

### *The number of rate/measurements*

- Multiple rater/measurement: If the experimental design of the reliability study involves three raters or more, then the "mean of k-raters" type should be selected.
- Single rater/measurement: If the experimental design of the reliability study involves one rater, then, "single rater" type should be selected.[4]

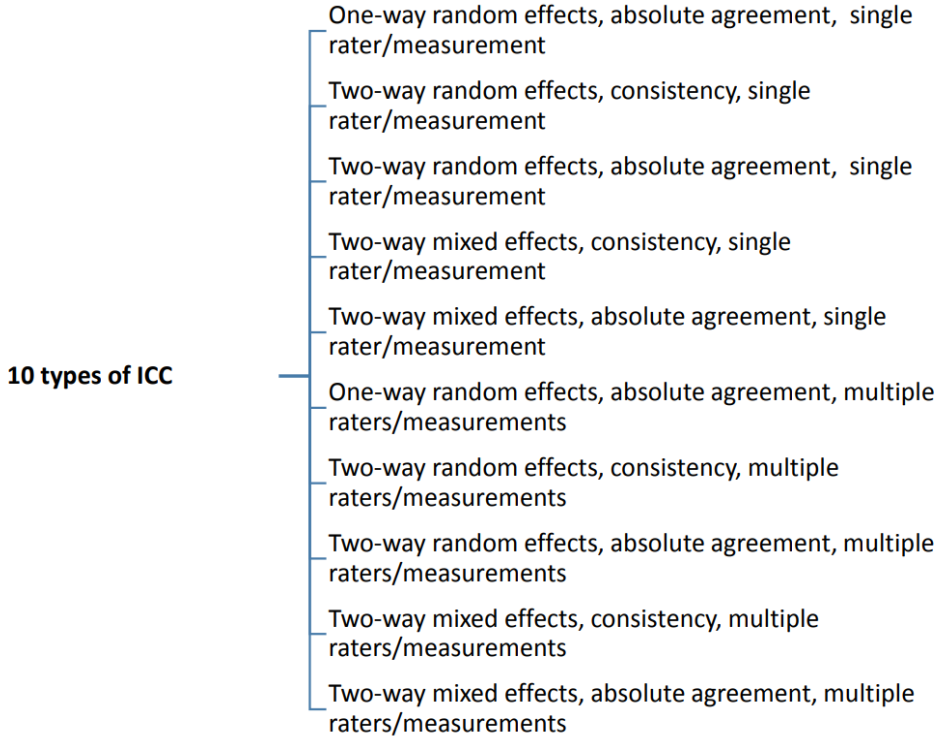


Figure 1. All types of ICC according to the models, type, Number of rater/measurement

### *Type*

- The Absolute agreement concerns the similarity of the values assigned to the same subjects by different raters ".[4, 12]
- Consistency concerns the assigned values to the same group of subjects correlated in an additive "manner".[4, 12]

### *ICC interpretation*

Literature review showed no standard values are presented for the interpretation of the acceptable reliability using ICC.[4] According to role of thumb, the ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.[13] Elsewhere, it is suggested that values less than 0.40 = poor, between 0.40 and 0.59 = fair, between 0.60 and 0.74 = good, and between 0.75 and 1.00 = excellent.[14]

Eventually, since there are different types of reliability, researchers must choose the appropriate method with sufficient knowledge. Also, because the calculation criteria have different types, the appropriate criterion should be selected according to the research conditions and objectives. The research results may be misleading if any of the reliability types and calculation criteria types are chosen incorrectly. Therefore, to make the results of the study more accurate and valuable. Medical researchers must seek help from relevant guidelines such as this study before conducting reliability analysis.

## References

1. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012;22(3):276-82.
2. Drost EA. Validity and reliability in social science research. Education Research and perspectives. 2011;38(1):105.
3. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? Physiotherapy. 2000;86(2):94-9.
4. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine. 2016;15(2):155-63.
5. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology. 2012;8(1):23.
6. Siegel S. Nonparametric statistics for the behavioral sciences. 1956.
7. Metrology JCfGi. Evaluation of measurement data—Guide to the expression of uncertainty in measurement. JCGM. 2008;100(2008):1-116.
8. Polit DF. Getting serious about test–retest reliability: a critique of retest research and some recommendations. Quality of Life Research. 2014;23(6):1713-20.
9. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.
10. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971;76(5):378-82.
11. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979;86(2):420-28.
12. McGraw KO, Wong SP. Forming inferences about some intraclass correlation

coefficients. Psychological methods. 1996;1(1):30-46.
13. Portney LG, Watkins MP. Foundations of clinical research: applications to practice: Pearson/Prentice Hall Upper Saddle River, NJ; 2009.
14. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment. 1994;6(4):284-90.