ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Semester Paper March 2025

Atreya Choudhury

Reliability

Submission Date: 31st March 2025

Adviser: Dr. Markus Kalisch

Abstract

We discuss various types of Intraclass Correlation Coefficients (ICC) as measures of reliability. Through practical examples, we guide the reader in selecting the most suitable ICC for their specific research question. Finally, we explore how different reliability questions can be addressed by appropriately choosing the ICC, using example datasets to illustrate the process.

Table of contents

Chapter 1

Introduction

Reliability is the study of the consistency of measurements. If we measure the same thing multiple times under the same conditions, we expect similar measurements. While several methods have been used in the past to study reliability, this document focuses on intraclass correlation coefficients (ICC) which help us understand how much of the variation is due to differences amongst subjects, differences amongst raters and measurement errors.

There are various questions one can ask while studying reliability. The questions influence the models chosen and consequently the ICCs. This document will primarily focus on the different factors affecting the choice of ICC and help readers find the quickest way to their required ICC. We focus on examples, the types of questions one can ask, and the corresponding ICCs. The theory behind the computation of different ICCs and the specifications of the models behind them are not discussed in this document and can be found in Mcgraw and Wong (1996) which is the primary reference for the discussion below.

1.1 Types of Reliability

- **Inter-rater reliability** measures agreement amongst different raters evaluating the same subjects. Consider the problem of radiologists examining X-ray images for multiple patients. If they consistently agree on whether each patient has a fracture, inter-rater reliability would be high.
- **Test-retest reliability** measures consistency of a test over time. If an IQ test is offered to a fixed group of subjects over different days, the test would have high test-retest reliability if the scores are similar over different days.
- **Intra-rater reliability** measures consistency of a rater over time. If a judge rates the performance of multiple gymnasts once on every day of the week, intra-rater reliability would be high if the gymnasts have consistent measurements over the different days.

There is subtle distinction between test-retest reliability and intra-rater reliability where the test-retest reliability is more concerned with the reliability of a test while assuming the rater effects are neglectable.

Chapter 2

Choosing the right ICC

There are three main factors affecting the choice of ICC - model, type and definition. We will describe what each mean and how to choose them appropriately.

2.1 Model

Model refers to the design choice of the statistical model used.

2.1.1 One-Way Random Effects

$$x_{ij} = \mu + r_i + w_{ij}$$

where μ (the population mean) is fixed, r_i (the row effects) are random, w_{ij} (the residual effects) are random, $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

In one way models, we do not model the effect of raters as a separate covariate. This means that the model assumes that there is no systematic way in which raters are assigned to subjects. This is quite rare as in most settings, we have the same set of raters assigned to all subjects.

However, there are setups which prohibit the same set of raters for all subjects. Consider multi-center studies where one set of raters assess a subgroup of subjects while another set of raters assess another subgroup. This is also the case when working with unmatched data where measurements are taken under unique conditions (which are random) that do not repeat for each subject. In such cases, one way models are appropriate.

2.1.2 Two-Way Random Effects

with interaction

$$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$$

and without interaction

$$x_{ij} = \mu + r_i + c_j + e_{ij}$$

where μ (the population mean) is fixed, r_i (the row effects) are random, c_j (the column effects) are random, rc_{ij} (the interaction effects) are random, e_{ij} (the residual effects) are random, $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

In contrast to one way models, two way models are used when raters are systematically assigned to subjects. This model assumes that raters are selected from a large population. This model should be used when we want to generalise our reliability study to describe any raters which come from the same population of raters chosen in the study.

2.1.3 Two-Way Mixed Effects

with interaction

$$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$$

and without interaction

$$x_{ij} = \mu + r_i + c_j + e_{ij}$$

where μ (the population mean) is fixed, r_i (the row effects) are random, c_j (the column effects) are fixed, rc_{ij} (the interaction effects) are random, e_{ij} (the residual effects) are random, $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

This model does not generalise for raters and should be used when the raters selected in the study are the only raters of interest.

2.2 Type

When assessing reliability using ICC, type selection refers to the choice between the measurement protocol of "single rater" (ICC$(\cdot, 1)$) and "mean of k raters" (ICC$(\cdot, k)$) (see convention in Table 2.1). Let us give a quick example to illustrate the difference.

Consider a study with 10 patients and 5 doctors where every doctor measures the post surgery pain level of every patient. To make some population level inferences, we use a two-way random effects model. Now, if we are asking the question, "If I randomly pick one doctor from a pool, how reliable would a single doctor's rating be?", we should use ICC$(\cdot, 1)$. However, if the questions is "If I randomly pick 5 doctors from a pool, how reliable would the mean of their ratings be?", we should use ICC$(\cdot, k)$ instead.

2.3 Definition

Another important question when assessing reliability is whether we consider absolute agreement (ICC$(A, \cdot)$) or consistency (ICC$(C, \cdot)$) (see convention in Table 2.1) amongst raters to be more important. Let us consider the same example from above with 10 patients and 5 doctors.

If we were interested in absolute agreement, we would be asking, "How closely do doctors agree on the exact score for each patient?". On the other hand, if we were interested in consistency, we would be asking, "how closely do doctors maintain the rank order of the different patients, even if their individual ratings differ?"

2.4 Additional Discussion

For two way models (random and mixed), we make a distinction between models with or without interaction effects. However, we disregard this difference for the computation of the ICC as we observe that for a given choice of type and definition, the ICC computed

is the same irrespective of the interaction effect. The exception to this is that the ICC is not estimable for fixed effect models with interactions when averaging scores (under consistency and absolute agreement).

Note that the convention (as seen in table 2.1) is different for one-way models. This is because for one-way models, we can only measure absolute agreement. Please note that the ICC calculation formulas are the same for random and mixed models in the two-way case. Hence, they share the same convention.

Table 2.1: Convention described in Mcgraw and Wong (1996)

Model	Type	Definition	Designation
one-way random effects	single rater	absolute agreement	ICC(1)
	mean of k raters	absolute agreement	ICC(k)
two-way random effects/ two-way mixed effects	single rater	absolute agreement	ICC($A,1$)
		consistency	ICC($C,1$)
	mean of k raters	absolute agreement	ICC(A,k)
		consistency	ICC(C,k)

2.5 Inference

The general guideline is that ICC values less than 0.5 indicate 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 0.9 indicate excellent reliability. When reporting ICC values in a study, it is important to report the reliability based on the confidence interval bounds of the computed ICC, rather than solely reporting the reliability of the raw ICC score itself.

Chapter 3

Examples

We now focus on case studies and choosing the right ICC to answer relevant questions. We use the `R` package `irr` (Gamer, Lemon, and Singh, 2019) for ICC computations.

```
suppressPackageStartupMessages({
    library(irr)
    library(kableExtra)
    library(lme4)
})
```

3.1 ethics dataset

We use the **ethics and student performance dataset** from Soto-Pérez, Ávila-Palet, and Núñez-Ríos (2022) which examines the relationship between various ethical ideologies of students and their academic performance. The study uses several instruments to measure ethical ideologies such as justice, utilitarianism, moral meaningfulness, and others (table 2 in Soto-Pérez et al. (2022)). Each latent variable is assessed using multiple instruments, with a 5-point Likert scale employed for each instrument. In this context, the students are the subjects, and the different instruments act as the raters.

For moral meaningfulness, four instruments are used. A key question here is whether the different instruments consistently agree in their assessment of various students. Let's begin by loading the data.

```
dep <- read.csv("data/DataEthicsPerformance.csv")
mm <- dep[, grep("^MM", names(dep))]
kbl(as.data.frame(apply(mm, 2, summary)))
```

The natural choice now would be a two-way mixed model. We choose a mixed model (fixed effects for the different instruments/tests) as we are only interested in studying the given instruments. We are interested in the individual ability of each test to gauge moral meaningfulness and the consistency of these tests. This implies an ICC$(C, 1)$.

```
icc(mm, model = "twoway", type = "consistency", unit = "single", r0 = 0.5)
```

```
 Single Score Intraclass Correlation
```

Table 3.1: Summary of moral meaningfulness instruments

	MM1	MM2	MM3	MM4
Min.	2.000000	2.000000	2.000000	2.000000
1st Qu.	4.000000	4.000000	4.750000	4.000000
Median	5.000000	5.000000	5.000000	5.000000
Mean	4.559702	4.548508	4.682836	4.679105
3rd Qu.	5.000000	5.000000	5.000000	5.000000
Max.	5.000000	5.000000	5.000000	5.000000

```
  Model: twoway
  Type : consistency

  Subjects = 268
    Raters = 4
  ICC(C,1) = 0.551

F-Test, H0: r0 = 0.5 ; H1: r0 > 0.5
F(267,801) = 1.18 , p = 0.0447

95%-Confidence Interval for ICC Population Values:
 0.492 < ICC < 0.609
```

At 95% confidence, we see evidence of poor to moderate inter-rater reliability (consistency amongst the different instruments). In contrast, if we are interested in the combined strength of the different tests (mean of k raters) and compute ICC(C, k), we get

```
icc(mm, model = "twoway", type = "consistency", unit = "average", r0 = 0.75)
```

```
Average Score Intraclass Correlation

  Model: twoway
  Type : consistency

  Subjects = 268
    Raters = 4
  ICC(C,4) = 0.831

F-Test, H0: r0 = 0.75 ; H1: r0 > 0.75
F(267,801) = 1.48 , p = 2.81e-05

95%-Confidence Interval for ICC Population Values:
 0.795 < ICC < 0.861
```

At 95% confidence, we see evidence of good average rater reliability (consistency of the mean of the different instruments).

We perform the same for the different latent variables (ethical ideologies).

```
eis <- c(
    "J" = "Justice", "R" = "Relativist", "E" = "Egoism",
```

Table 3.2: Summary of ICC(C, 1) of different ethical ideologies

ideology	instruments	ICC	lower.bound	upper.bound
Justice	2	0.5530814	0.4641638	0.6309628
Relativist	2	0.2301149	0.1136137	0.3403737
Egoism	2	0.5844004	0.4997067	0.6580259
Utilitarianism	3	0.5611770	0.4951321	0.6239296
Deontology	2	0.5073339	0.4127561	0.5910885
Moral Meaningfulness	4	0.5506959	0.4920485	0.6085068
Citizenship Behaviour	3	0.5162236	0.4469391	0.5828777
In-role Performance	2	0.3664833	0.2581734	0.4656951

```
    "U" = "Utilitarianism", "D" = "Deontology",
    "MM" = "Moral Meaningfulness", "CB" = "Citizenship Behaviour",
    "IRP" = "In-role Performance"
)
results <- data.frame()

for (ei in names(eis)) {
    ratings <- dep[, grep(paste0("^", ei), names(dep))]
    ratings.icc <- icc(
        ratings,
        model = "twoway", type = "consistency", unit = "single"
    )
    results <- rbind(results, data.frame(
        ideology = eis[[ei]],
        instruments = ratings.icc$raters,
        ICC = ratings.icc$value,
        lower.bound = ratings.icc$lbound,
        upper.bound = ratings.icc$ubound
    ))
}
kbl(results)
```

We see poor to moderate inter-rater reliability for most latent variables. Similar to above, we see a much stronger average rater reliability.

```
results <- data.frame()

for (ei in names(eis)) {
    ratings <- dep[, grep(paste0("^", ei), names(dep))]
    ratings.icc <- icc(
        ratings,
        model = "twoway", type = "consistency", unit = "average"
    )
    results <- rbind(results, data.frame(
        ideology = eis[[ei]],
        instruments = ratings.icc$raters,
        ICC = ratings.icc$value,
```

Table 3.3: Summary of ICC(C, k) of different ethical ideologies

ideology	instruments	ICC	lower.bound	upper.bound
Justice	2	0.7122375	0.6340326	0.7737305
Relativist	2	0.3741356	0.2040451	0.5078788
Egoism	2	0.7376928	0.6664059	0.7937462
Utilitarianism	3	0.7932376	0.7463312	0.8326984
Deontology	2	0.6731540	0.5843275	0.7429989
Moral Meaningfulness	4	0.8305847	0.7948620	0.8614439
Citizenship Behaviour	3	0.7619735	0.7079744	0.8074010
In-role Performance	2	0.5363890	0.4103940	0.6354597

```
        lower.bound = ratings.icc$lbound,
        upper.bound = ratings.icc$ubound
    ))
}
kbl(results)
```

3.2 movies dataset

The **MovieLens 1M** dataset (Harper and Konstan, 2015) dataset has 1M ratings from around 6k users and 4k movies. An interesting question here would be whether there is a consensus between raters in different movies.

To combat the sparsity of this dataset, we filter it for the most watched movies and find the raters who rated all of them. Firstly, we find all the users who have watched the 60 most rated movies.

```
movies.raw <- readLines("data/ratings.dat")
movies.raw <- gsub("::", ",", movies.raw)
movies <- read.csv(
    text = movies.raw,
    col.names = c("user", "movie", "rating", "timestamp")
)
movies$timestamp <- NULL
N <- 60

movie_counts <- table(movies$movie)
top_movies <- sort(movie_counts, decreasing = TRUE)[1:N]

movies <- movies[movies$movie %in% names(top_movies), ]
umt <- table(movies$user, movies$movie)

users_rated_all <- rownames(umt)[rowSums(umt > 0) == N]

length(users_rated_all)
```

```
[1] 6
```

Now, we filter those movies (and their ratings) which have been watched by these users and reshape the data for our purposes. We want to use a two-way random effects model. We are interested in the consistency of different raters as well as the consistency of the mean of the ratings. Hence, we compute ICC$(C, 1)$ and ICC(C, k).

```
movies <- movies[movies$user %in% users_rated_all, ]

movies.wide <- reshape(movies,
    idvar = "movie",
    timevar = "user",
    direction = "wide"
)
movies.wide$movie <- NULL

icc(movies.wide, model = "twoway", type = "consistency", unit = "single")
```

```
 Single Score Intraclass Correlation

   Model: twoway
   Type : consistency

   Subjects = 60
     Raters = 6
   ICC(C,1) = 0.265

 F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(59,295) = 3.17 , p = 5.77e-11

 95%-Confidence Interval for ICC Population Values:
  0.165 < ICC < 0.39
```

```
icc(movies.wide, model = "twoway", type = "consistency", unit = "average")
```

```
 Average Score Intraclass Correlation

   Model: twoway
   Type : consistency

   Subjects = 60
     Raters = 6
   ICC(C,6) = 0.684

 F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(59,295) = 3.17 , p = 5.77e-11

 95%-Confidence Interval for ICC Population Values:
  0.542 < ICC < 0.794
```

We observe poor inter-rater reliability and moderate to good average reliability.

3.3 sleepstudy dataset

The `sleepstudy` dataset, available in the `lme4` package, is from a sleep deprivation study in which the average reaction time of different subjects was measured over a ten-day period.

This is an example of a problem where we are interested in test-retest reliability. We start with a mixed effect model and compute ICC$(A, 1)$ to check agreement of measurements across different days.

```
data(sleepstudy)
sleepstudy.wide <- reshape(sleepstudy,
    idvar = "Subject",
    timevar = "Days",
    direction = "wide"
)
sleepstudy.wide$Subject <- NULL

icc(sleepstudy.wide, model = "twoway", type = "agreement", unit = "single")
```

```
 Single Score Intraclass Correlation

   Model: twoway
   Type : agreement

   Subjects = 18
     Raters = 10
   ICC(A,1) = 0.413

 F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(17,35.4) = 14.9 , p = 2.25e-11

 95%-Confidence Interval for ICC Population Values:
  0.228 < ICC < 0.644
```

As expected, we see a fairly wide confidence interval for the ICC indicating poor to moderate test-retest reliability.

3.4 anxiety dataset

The `anxiety` dataset, available in the `irr` package, contains anxiety ratings of 20 subjects, rated by 3 raters.

The most straightforward question here is related to inter-rater reliability, for which we compute ICC$(C, 1)$.

```
data(anxiety)
icc(anxiety, model = "twoway", type = "consistency", unit = "single")
```

```
 Single Score Intraclass Correlation

   Model: twoway
   Type : consistency
```

```
   Subjects = 20
     Raters = 3
   ICC(C,1) = 0.216

 F-Test, H0: r0 = 0 ; H1: r0 > 0
   F(19,38) = 1.83 , p = 0.0562

 95%-Confidence Interval for ICC Population Values:
  -0.046 < ICC < 0.522
```

We observe poor to moderate inter-rater reliability which improves slightly when considering the reliability of the average.

```
icc(anxiety, model = "twoway", type = "consistency", unit = "average")
```

```
 Average Score Intraclass Correlation

   Model: twoway
   Type : consistency

   Subjects = 20
     Raters = 3
   ICC(C,3) = 0.453

 F-Test, H0: r0 = 0 ; H1: r0 > 0
   F(19,38) = 1.83 , p = 0.0562

 95%-Confidence Interval for ICC Population Values:
  -0.153 < ICC < 0.766
```

References

Gamer, M., J. Lemon, and I. F. P. Singh (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Harper, F. M. and J. A. Konstan (2015, December). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5*(4).

Mcgraw, K. and S. Wong (1996, 03). Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*, 30–46.

Soto-Pérez, M., J.-E. Ávila-Palet, and J. E. Núñez-Ríos (2022). Justice, deontology and moral meaningfulness as factors to improve student performance and academic achievement. *Journal of Academic Ethics 20*(3), 375–397.

Epilogue

I would like to thank Dr. Markus Kalisch for his time and guidance.