# Estimating Interrater Reliability from Planned-Missing Data

## Demonstration of the ICC4IRR application

Debby ten Hove

Department of Education and Family Studies, Vrije Universiteit Amsterdam
*D.ten.Hove@VU.nl*

August 26, 2025

**Contributors**

Dr. Letty Koopman
University of Groningen

Tasos Psychogiopoulous, MSc.
University of Amsterdam

**ICC4IRR Application**

https://tasospsy.shinyapps.io/icc4irr_app/

**Example Data**

https://github.com/icc4irr/app/blob/main/sample-data/Example_ratings_EARLI2025.csv

# Motivating example: Teaching Quality

Education researchers use **observation instruments** to evaluate teaching quality

Example: **Educational inspectors** *(raters)* evaluate **teaching skills** *(attribute)* of **teachers** *(subjects)*

# Motivating example: Teaching Quality

Such ratings are used in **practice** and in **research**

# Motivating example: Teaching Quality

Such ratings are used in **practice** and in **research**

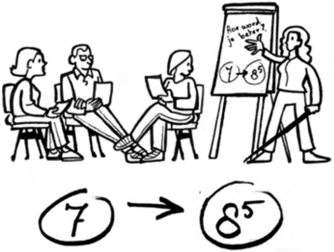
**In practice** to make decision about teachers or schools.

# Motivating example: Teaching Quality

Such ratings are used in **practice** and in **research**

**In practice** to make decision about teachers or schools.

**In research** to study differences across teachers or schools.
For example in intervention studies.



# Motivating example: Teaching Quality

**Important** that (the variation in) observed scores originate in differences across teachers or schools, and as little as possible in rater effects.

| **Use of scores** | **Problems due to rater effects (noise)** |
|---|---|
| Regression techniques | **(Attenuation) Bias and loss of precision** |
| Decisions about individuals | **Incorrect decisions** |

→ Important to investigate the **interrater reliability** (IRR).

# Intraclass correlation coefficients

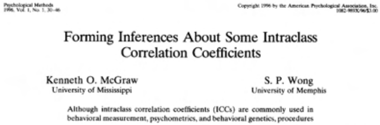
## Intraclass correlation coefficients (ICCs) for Interrater reliability (IRR)

To which degree can we differentiate between subjects, hence generalize subject scores over raters?

(Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979)

- Applicable to quantitative data
- Rooted in Generalizability theory (Cronbach et al., 1963)
- Coefficients for absolute and relative decision making
- Available for $\geq 2$ raters

# Intraclass correlation coefficients



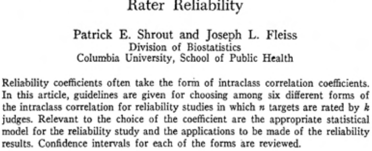
Psychological Methods
1996, Vol. 1, No. 1, 30–46

## Forming Inferences About Some Intraclass Correlation Coefficients

Kenneth O. McGraw
University of Mississippi

S. P. Wong
University of Memphis

Although intraclass correlation coefficients (ICCs) are commonly used in behavioral measurement, psychometrics, and behavioral genetics, procedures

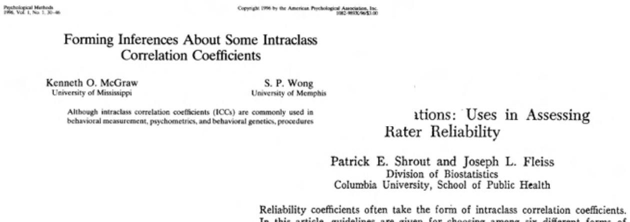


## ations: Uses in Assessing Rater Reliability

Patrick E. Shrout and Joseph L. Fleiss
Division of Biostatistics
Columbia University, School of Public Health

Reliability coefficients often take the form of intraclass correlation coefficients. In this article, guidelines are given for choosing among six different forms of the intraclass correlation for reliability studies in which *n* targets are rated by *k* judges. Relevant to the choice of the coefficient are the appropriate statistical model for the reliability study and the applications to be made of the reliability results. Confidence intervals for each of the forms are reviewed.

# Intraclass correlation coefficients



Psychological Methods
1996, Vol. 1, No. 1, 30–46

Copyright 1996 by the American Psychological Association, Inc.
1082-989X/96/$3.00

Forming Inferences About Some Intraclass Correlation Coefficients

Kenneth O. McGraw
University of Mississippi

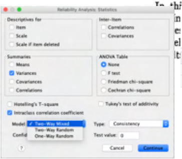
S. P. Wong
University of Memphis

Although intraclass correlation coefficients (ICCs) are commonly used in behavioral measurement, psychometrics, and behavioral genetics, procedures

ıtions: Uses in Assessing Rater Reliability

Patrick E. Shrout and Joseph L. Fleiss
Division of Biostatistics
Columbia University, School of Public Health

Reliability coefficients often take the form of intraclass correlation coefficients.



Reliability Analysis: Statistics

Descriptives for
□ Item
□ Scale
□ Scale if item deleted

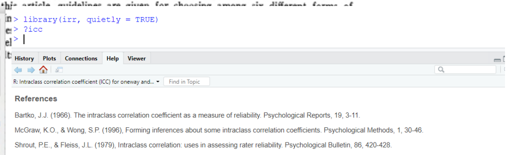
Inter-Item
□ Correlations
□ Covariances

Summaries
□ Means
☑ Variances
□ Covariances
□ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

□ Hotelling's T-square
□ Tukey's test of additivity
☑ Intraclass correlation coefficient

Model: ✓ Two-Way Mixed / Two-Way Random / One-Way Random
Type: Consistency
Test value: 0



```
> library(irr, quietly = TRUE)
> ?icc
> 
```

R: Intraclass correlation coefficient (ICC) for oneway and...

References

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

McGraw, K.O., & Wong, S.P. (1996), Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30-46.

Shrout, P.E., & Fleiss, J.L. (1979), Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

# ICCs for Interrater Reliability

## Two-Way Data

- Each subject is assessed by the same $k \geq 2$ raters.

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

# ICCs for Interrater Reliability

## Two-Way Data

- Each subject is assessed by the same $k \geq 2$ raters.

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr.e}^2$$

# ICCs for Interrater Reliability

## Two-Way Data

- Each subject is assessed by the same $k \geq 2$ raters.

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr.e}^2$$

## Interrater Agreement

$$ICC(A,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr.e}^2}{k}}$$

# ICCs for Interrater Reliability

## Two-Way Data

- Each subject is assessed by the same $k \geq 2$ raters.

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr.e}^2$$

## Interrater Agreement

$$ICC(A,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr.e}^2}{k}}$$

## Interrater Consistency

$$ICC(C,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{sr.e}^2}{k}}$$

# Practice: Planned-Missing Data

## Complete data

- ICC definitions *and* estimation methods require complete data

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

# Practice: Planned-Missing Data

## Complete data

- ICC definitions *and* estimation methods require complete data

| Subject | Rater | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | $y_{43}$ |
| 5 | $y_{51}$ | $y_{52}$ | $y_{53}$ |
| 6 | $y_{61}$ | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | $y_{73}$ |
| 8 | $y_{81}$ | $y_{82}$ | $y_{83}$ |
| 9 | $y_{91}$ | $y_{92}$ | $y_{93}$ |

## Incomplete data

- Most educational studies use a planned-missing design. For example:

| Teacher | Educational Inspector | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | $y_{11}$ | $y_{12}$ | – |
| 2 | $y_{21}$ | – | $y_{23}$ |
| 3 | – | $y_{32}$ | $y_{33}$ |
| 4 | $y_{41}$ | $y_{42}$ | – |
| 5 | $y_{51}$ | – | $y_{53}$ |
| 6 | – | $y_{62}$ | $y_{63}$ |
| 7 | $y_{71}$ | $y_{72}$ | – |
| 8 | $y_{81}$ | – | $y_{83}$ |
| 9 | – | $y_{92}$ | $y_{93}$ |

# ICCs for planned-missing data

**Ten Hove, Jorgensen and Van der Ark (2024). Updated Guidelines on Selecting ICCs for IRR.**

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2$$

# ICCs for planned-missing data

Ten Hove, Jorgensen and Van der Ark (2024). Updated Guidelines on Selecting ICCs for IRR.

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2$$

## Interrater Agreement

$$ICC(A,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\boldsymbol{\sigma}_{r}^{2} + \sigma_{sr}^2}{k}}$$

Account for unbalanced number of raters:

$$ICC(A,\hat{k}) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr}^2}{\hat{k}}}$$

# ICCs for planned-missing data

Ten Hove, Jorgensen and Van der Ark (2024). Updated Guidelines on Selecting ICCs for IRR.

## Variance Decomposition

$$\sigma_y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2$$

## Interrater Agreement

$$ICC(A,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr}^2}{k}}$$

Account for unbalanced number of raters:

$$ICC(A,\hat{k}) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr}^2}{\hat{k}}}$$

## Interrater Consistency

$$ICC(C,k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{sr}^2}{k}}$$

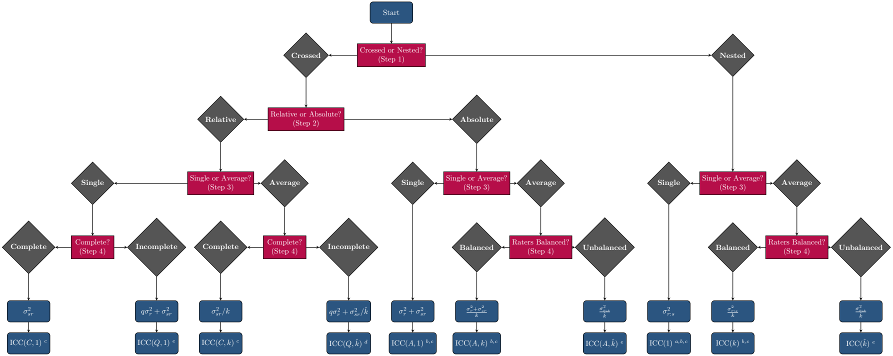
Account for partial non-overlapping raters:

$$ICC(Q,\hat{k}) = \frac{\sigma_s^2}{\sigma_s^2 + q * \sigma_r^2 + \frac{\sigma_{sr}^2}{\hat{k}}}$$

$$\hat{k} = \left(\frac{k_1^{-1} + k_2^{-1} + \ldots + k_l^{-1}}{N_s}\right)^{-1}; q = \frac{1}{\hat{k}} - \frac{\sum_i \sum_{i'} \frac{k_{i,i'}}{k_i k_{i'}}}{N_s(N_s - 1)}$$

# Overview ICC Selection

See: Ten Hove, Jorgensen and Van der Ark (2024). Updated Guidelines on Selecting an ICC for IRR.



Start

Crossed or Nested? (Step 1)

Crossed

Nested

Relative or Absolute? (Step 2)

Relative

Absolute

Single or Average? (Step 3)

Single

Average

Single or Average? (Step 3)

Single

Average

Single or Average? (Step 3)

Single

Average

Complete? (Step 4)

Complete

Incomplete

Complete? (Step 4)

Complete

Incomplete

Raters Balanced? (Step 4)

Balanced

Unbalanced

Raters Balanced? (Step 4)

Balanced

Unbalanced

$\sigma^2_{sr}$

ICC$(C,1)$ $^{c}$

$q\sigma^2_r + \sigma^2_{sr}$

ICC$(Q,1)$ $^{e}$

$\sigma^2_{sr}/k$

ICC$(C,k)$ $^{c}$

$q\sigma^2_r + \sigma^2_{sr}/\hat{k}$

ICC$(Q,\hat{k})$ $^{d}$

$\sigma^2_r + \sigma^2_{sr}$

ICC$(A,1)$ $^{b,c}$

$\frac{\sigma^2_r+\sigma^2_{sr}}{k}$

ICC$(A,k)$ $^{b,c}$

$\frac{\sigma^2_{r:s}}{\hat{k}}$

ICC$(A,\hat{k})$ $^{e}$

$\sigma^2_{r:s}$

ICC$(1)$ $^{a,b,c}$

$\frac{\sigma^2_{r:s}}{k}$

ICC$(k)$ $^{b,c}$

$\frac{\sigma^2_{r:s}}{\hat{k}}$

ICC$(\hat{k})$ $^{e}$

# Estimating ICCs from Planned-Missing Data

## Compared three estimation methods for ICCs:

- **MCMC**: Markov chain Monte Carlo Estimation of hierarchical models (LoPilato et al., 2015; Ten Hove et al., 2020, 2021)
- **MLE-R**: Maximum likelihood estimation of **R**andom effects models (Marcoulides, 1990; cf. Jiang, 2018; Ten Hove et al., 2021)
- **MLE-CF**: Maximum likelihood estimation of **C**ommon-**F**actor models (Jorgensen, 2021; Marcoulides, 1996; Vispoel et al., 2018a, 2019))

For various design factors (e.g., $K, N, \hat{k}, \sigma^2$)

## Based on (*among other things*):

- Computational accuracy
- 95% (B)CI coverage rates

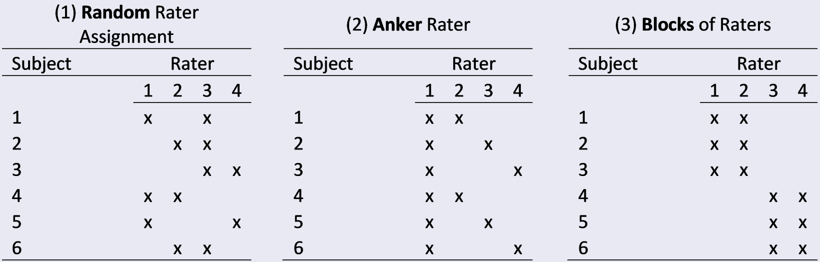
SCAN ME

# Estimating ICCs from Planned-Missing Data

## AND what if..

A different observational design were used?

## Examples of Planned Missing Designs



(1) **Random** Rater Assignment

| Subject | Rater | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | x | | x | |
| 2 | | x | x | |
| 3 | | | x | x |
| 4 | x | x | | |
| 5 | x | | | x |
| 6 | | x | x | |

(2) **Anker** Rater

| Subject | Rater | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | x | x | | |
| 2 | x | | x | |
| 3 | x | | | x |
| 4 | x | x | | |
| 5 | x | | x | |
| 6 | x | | | x |

(3) **Blocks** of Raters

| Subject | Rater | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | x | x | | |
| 2 | x | x | | |
| 3 | x | x | | |
| 4 | | | x | x |
| 5 | | | x | x |
| 6 | | | x | x |

# Conclusions

- **Simulation 1: MLE of random effects models** useful for estimating ICCs for IRR from typical observational studies
  - *Also* most **User-friendly** because it converges in most conditions and only takes seconds.
- **Simulation 2:** Type of (planned missing) design does not matter much with respect to (SE) bias and coverage of ICCs

Conclusion: **MLE of random effects models** very useful for (interrater)reliability studies.

# Current Work: ICC4IRR application

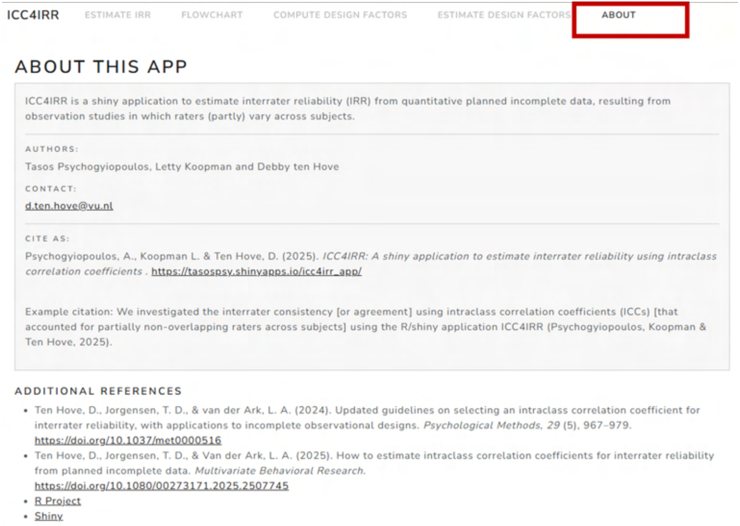
- **Under Development**: Shiny app to estimate ICCs for IRR

  Psychogiopoulos, Koopman & Ten Hove (2025)

- **In progress**: Tutorial and guidance in planning rater studies

SCAN ME

# ICC4IRR application



## ABOUT THIS APP

ICC4IRR is a shiny application to estimate interrater reliability (IRR) from quantitative planned incomplete data, resulting from observation studies in which raters (partly) vary across subjects.

AUTHORS:

Tasos Psychogyiopoulos, Letty Koopman and Debby ten Hove

CONTACT:

d.ten.hove@vu.nl

CITE AS:

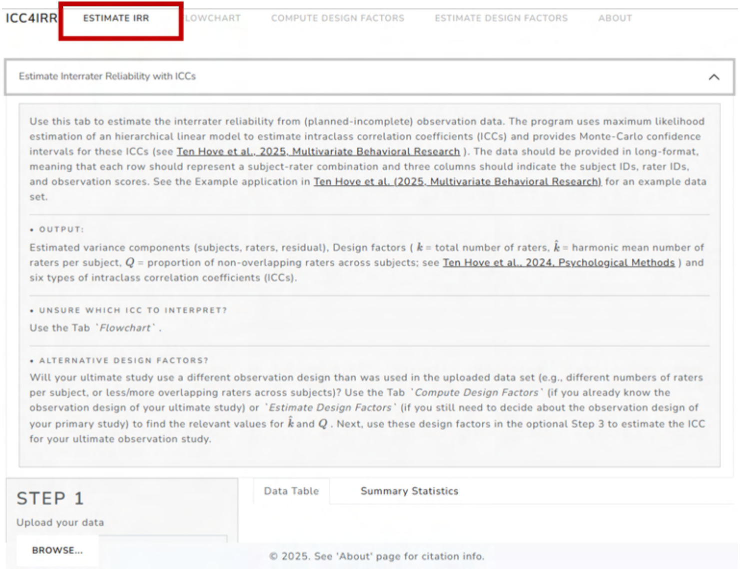
Psychogyiopoulos, A., Koopman L. & Ten Hove, D. (2025). *ICC4IRR: A shiny application to estimate interrater reliability using intraclass correlation coefficients* . https://tasospsy.shinyapps.io/icc4irr_app/

Example citation: We investigated the interrater consistency [or agreement] using intraclass correlation coefficients (ICCs) [that accounted for partially non-overlapping raters across subjects] using the R/shiny application ICC4IRR (Psychogyiopoulos, Koopman & Ten Hove, 2025).

### ADDITIONAL REFERENCES

- Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods, 29* (5), 967–979. https://doi.org/10.1037/met0000516
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2025). How to estimate intraclass correlation coefficients for interrater reliability from planned incomplete data. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2025.2507745
- R Project
- Shiny

# ICC4IRR application



Estimate Interrater Reliability with ICCs

Use this tab to estimate the interrater reliability from (planned-incomplete) observation data. The program uses maximum likelihood estimation of an hierarchical linear model to estimate intraclass correlation coefficients (ICCs) and provides Monte-Carlo confidence intervals for these ICCs (see Ten Hove et al., 2025, Multivariate Behavioral Research ). The data should be provided in long-format, meaning that each row should represent a subject-rater combination and three columns should indicate the subject IDs, rater IDs, and observation scores. See the Example application in Ten Hove et al. (2025, Multivariate Behavioral Research) for an example data set.

- OUTPUT:

Estimated variance components (subjects, raters, residual), Design factors ( $k$ = total number of raters, $\hat{k}$ = harmonic mean number of raters per subject, $Q$ = proportion of non-overlapping raters across subjects; see Ten Hove et al., 2024, Psychological Methods ) and six types of intraclass correlation coefficients (ICCs).

- UNSURE WHICH ICC TO INTERPRET?

Use the Tab `*Flowchart*` .

- ALTERNATIVE DESIGN FACTORS?

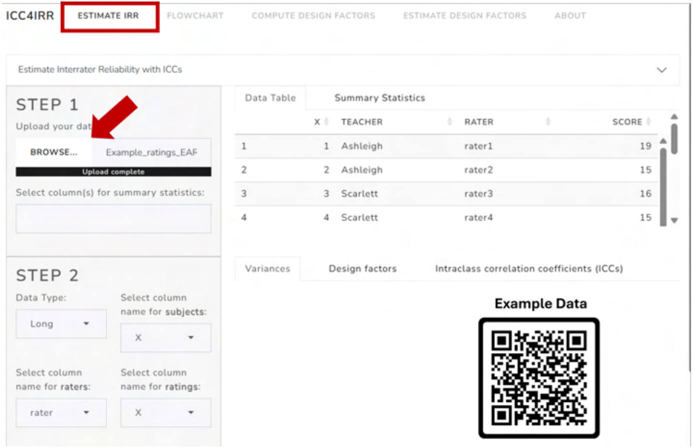
Will your ultimate study use a different observation design than was used in the uploaded data set (e.g., different numbers of raters per subject, or less/more overlapping raters across subjects)? Use the Tab `*Compute Design Factors*` (if you already know the observation design of your ultimate study) or `*Estimate Design Factors*` (if you still need to decide about the observation design of your primary study) to find the relevant values for $\hat{k}$ and $Q$ . Next, use these design factors in the optional Step 3 to estimate the ICC for your ultimate observation study.

STEP 1

Upload your data

© 2025. See 'About' page for citation info.

# ICC4IRR application



Estimate Interrater Reliability with ICCs

STEP 1

Upload your dat

Example_ratings_EAF

Upload complete

Select column(s) for summary statistics:

STEP 2

Data Type: Long

Select column name for subjects: X

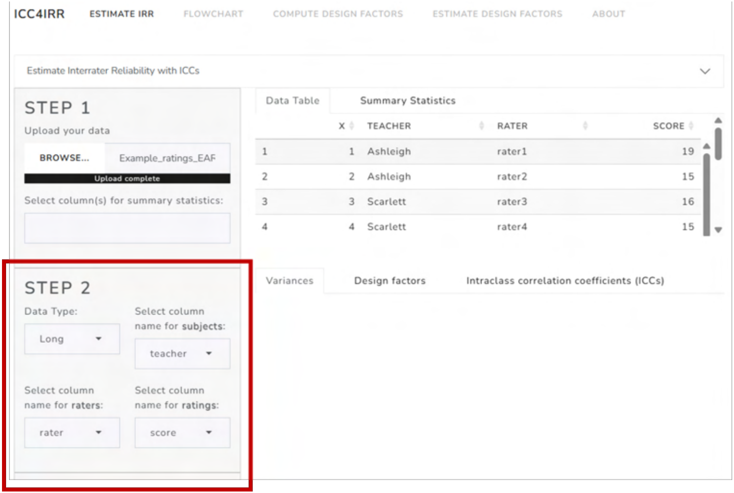
Select column name for raters: rater

Select column name for ratings: X

| | X | TEACHER | RATER | SCORE |
|---|---|---|---|---|
| 1 | 1 | Ashleigh | rater1 | 19 |
| 2 | 2 | Ashleigh | rater2 | 15 |
| 3 | 3 | Scarlett | rater3 | 16 |
| 4 | 4 | Scarlett | rater4 | 15 |

Example Data

# ICC4IRR application



Estimate Interrater Reliability with ICCs

STEP 1

Upload your data

Example_ratings_EAF

Select column(s) for summary statistics:

STEP 2

Data Type: Long

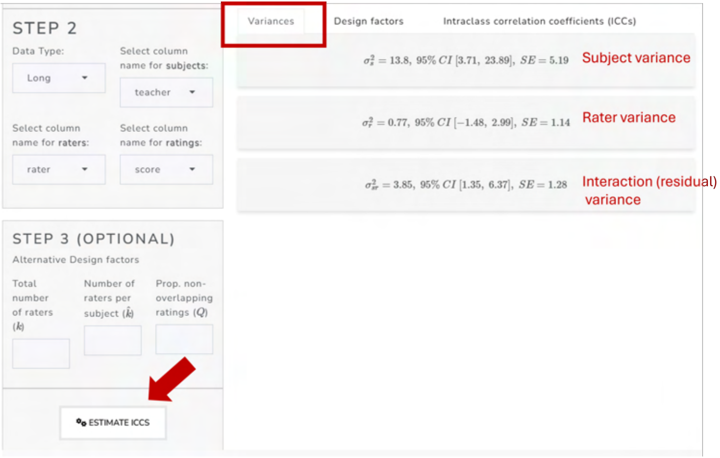
Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

| | X | TEACHER | RATER | SCORE |
|---|---|---|---|---|
| 1 | 1 | Ashleigh | rater1 | 19 |
| 2 | 2 | Ashleigh | rater2 | 15 |
| 3 | 3 | Scarlett | rater3 | 16 |
| 4 | 4 | Scarlett | rater4 | 15 |

# ICC4IRR application



STEP 2

Data Type: Long

Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

STEP 3 (OPTIONAL)

Alternative Design factors

Total number of raters ($k$):

Number of raters per subject ($\hat{k}$):

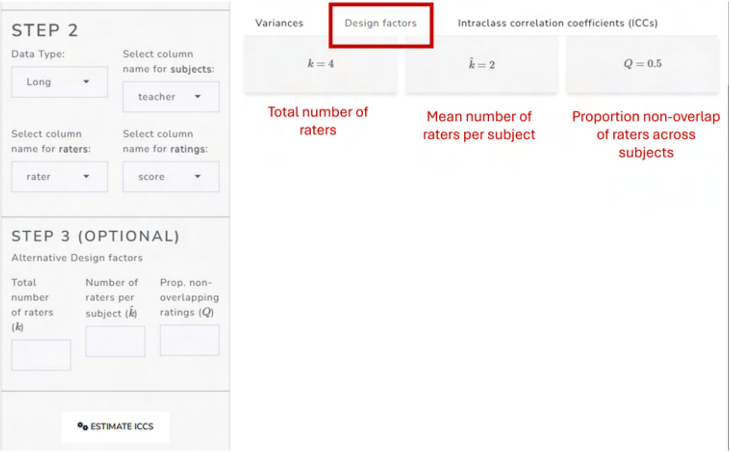
Prop. non-overlapping ratings ($Q$):

$\sigma_s^2 = 13.8,\ 95\%\ CI\ [3.71,\ 23.89],\ SE = 5.19$ Subject variance

$\sigma_r^2 = 0.77,\ 95\%\ CI\ [-1.48,\ 2.99],\ SE = 1.14$ Rater variance

$\sigma_{sr}^2 = 3.85,\ 95\%\ CI\ [1.35,\ 6.37],\ SE = 1.28$ Interaction (residual) variance

# ICC4IRR application



## STEP 2

Data Type: Long

Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

## STEP 3 (OPTIONAL)

Alternative Design factors

Total number of raters ($k$):

Number of raters per subject ($\hat{k}$):

Prop. non-overlapping ratings ($Q$):

$k = 4$

$\hat{k} = 2$

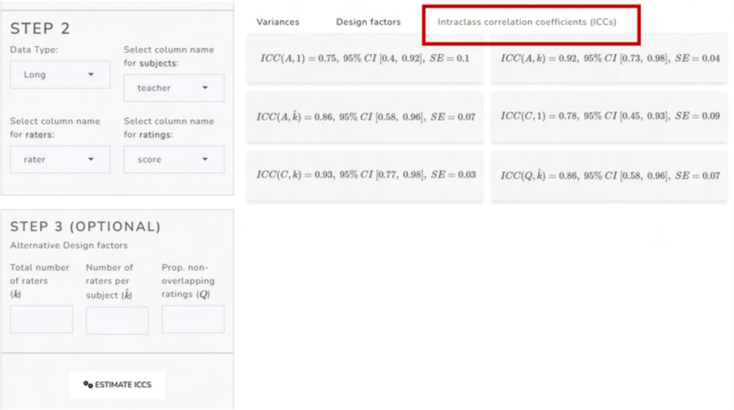
$Q = 0.5$

Total number of raters

Mean number of raters per subject

Proportion non-overlap of raters across subjects

# ICC4IRR application



STEP 2

Data Type: Long

Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

STEP 3 (OPTIONAL)

Alternative Design factors

Total number of raters ($k$)

Number of raters per subject ($\hat{k}$)

Prop. non-overlapping ratings ($Q$)

$ICC(A,1) = 0.75,\ 95\%\,CI\ [0.4,\ 0.92],\ SE = 0.1$

$ICC(A,k) = 0.92,\ 95\%\,CI\ [0.73,\ 0.98],\ SE = 0.04$

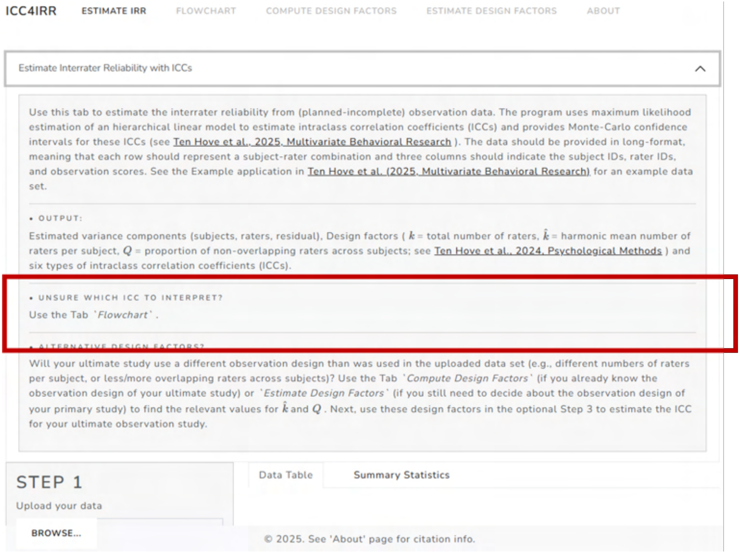
$ICC(A,\hat{k}) = 0.86,\ 95\%\,CI\ [0.58,\ 0.96],\ SE = 0.07$

$ICC(C,1) = 0.78,\ 95\%\,CI\ [0.45,\ 0.93],\ SE = 0.09$

$ICC(C,k) = 0.93,\ 95\%\,CI\ [0.77,\ 0.98],\ SE = 0.03$

$ICC(Q,\hat{k}) = 0.86,\ 95\%\,CI\ [0.58,\ 0.96],\ SE = 0.07$

# ICC4IRR application



Estimate Interrater Reliability with ICCs

Use this tab to estimate the interrater reliability from (planned-incomplete) observation data. The program uses maximum likelihood estimation of an hierarchical linear model to estimate intraclass correlation coefficients (ICCs) and provides Monte-Carlo confidence intervals for these ICCs (see Ten Hove et al., 2025, Multivariate Behavioral Research ). The data should be provided in long-format, meaning that each row should represent a subject-rater combination and three columns should indicate the subject IDs, rater IDs, and observation scores. See the Example application in Ten Hove et al. (2025, Multivariate Behavioral Research) for an example data set.

- OUTPUT:

Estimated variance components (subjects, raters, residual), Design factors ( $k$ = total number of raters, $\hat{k}$ = harmonic mean number of raters per subject, $Q$ = proportion of non-overlapping raters across subjects; see Ten Hove et al., 2024, Psychological Methods ) and six types of intraclass correlation coefficients (ICCs).

- UNSURE WHICH ICC TO INTERPRET?

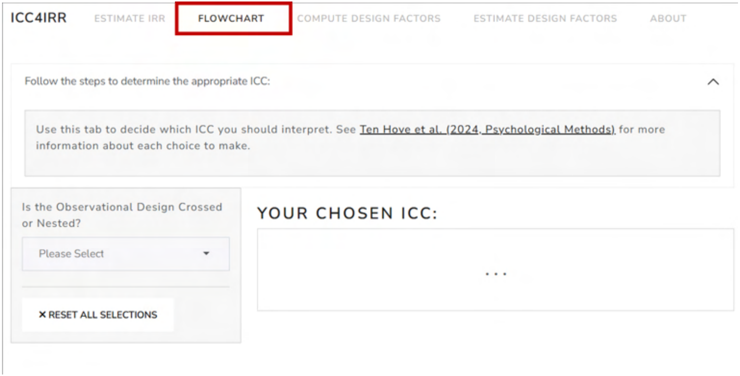
Use the Tab `*Flowchart*` .

- ALTERNATIVE DESIGN FACTORS?

Will your ultimate study use a different observation design than was used in the uploaded data set (e.g., different numbers of raters per subject, or less/more overlapping raters across subjects)? Use the Tab `*Compute Design Factors*` (if you already know the observation design of your ultimate study) or `*Estimate Design Factors*` (if you still need to decide about the observation design of your primary study) to find the relevant values for $\hat{k}$ and $Q$ . Next, use these design factors in the optional Step 3 to estimate the ICC for your ultimate observation study.

© 2025. See 'About' page for citation info.

# ICC4IRR application



Follow the steps to determine the appropriate ICC:

Use this tab to decide which ICC you should interpret. See Ten Hove et al. (2024, Psychological Methods) for more information about each choice to make.

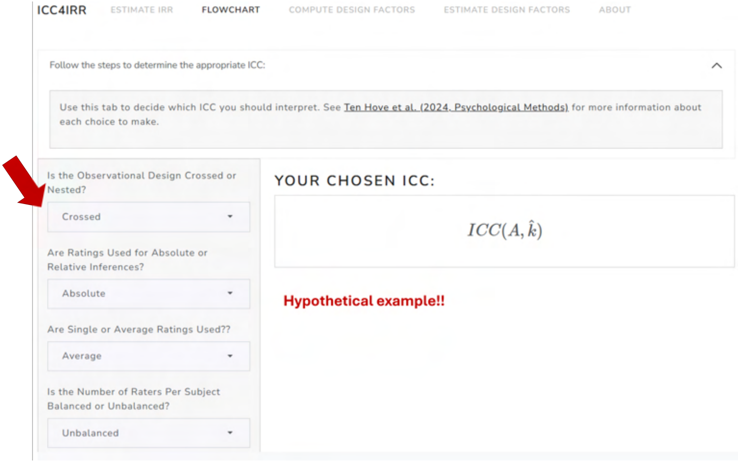
Is the Observational Design Crossed or Nested?

Please Select

YOUR CHOSEN ICC:

...

# ICC4IRR application



Follow the steps to determine the appropriate ICC:

Use this tab to decide which ICC you should interpret. See Ten Hove et al. (2024, Psychological Methods) for more information about each choice to make.

Is the Observational Design Crossed or Nested?
Crossed

Are Ratings Used for Absolute or Relative Inferences?
Absolute

Are Single or Average Ratings Used??
Average

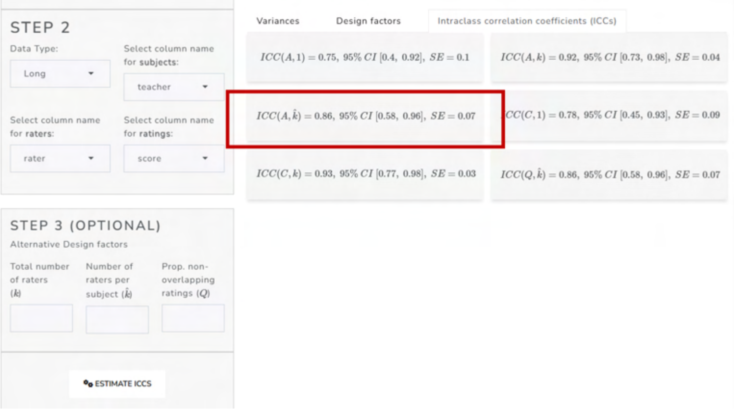
Is the Number of Raters Per Subject Balanced or Unbalanced?
Unbalanced

YOUR CHOSEN ICC:

$ICC(A, \hat{k})$

**Hypothetical example!!**

# ICC4IRR application



STEP 2

Data Type: Long

Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

STEP 3 (OPTIONAL)

Alternative Design factors

Total number of raters ($k$):

Number of raters per subject ($\hat{k}$):

Prop. non-overlapping ratings ($Q$):

$ICC(A,1) = 0.75,\ 95\%\ CI\ [0.4,\ 0.92],\ SE = 0.1$

$ICC(A,k) = 0.92,\ 95\%\ CI\ [0.73,\ 0.98],\ SE = 0.04$

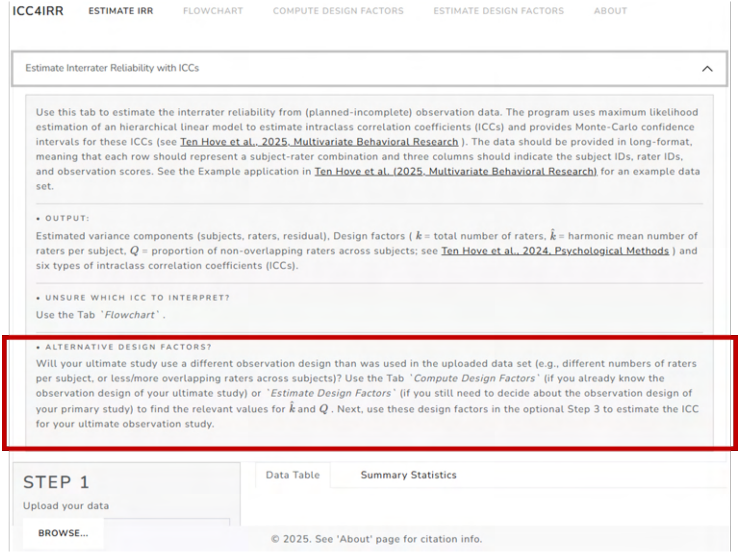
$ICC(A,\hat{k}) = 0.86,\ 95\%\ CI\ [0.58,\ 0.96],\ SE = 0.07$

$ICC(C,1) = 0.78,\ 95\%\ CI\ [0.45,\ 0.93],\ SE = 0.09$

$ICC(C,k) = 0.93,\ 95\%\ CI\ [0.77,\ 0.98],\ SE = 0.03$

$ICC(Q,\hat{k}) = 0.86,\ 95\%\ CI\ [0.58,\ 0.96],\ SE = 0.07$

# ICC4IRR application



Use this tab to estimate the interrater reliability from (planned-incomplete) observation data. The program uses maximum likelihood estimation of an hierarchical linear model to estimate intraclass correlation coefficients (ICCs) and provides Monte-Carlo confidence intervals for these ICCs (see Ten Hove et al., 2025, Multivariate Behavioral Research ). The data should be provided in long-format, meaning that each row should represent a subject-rater combination and three columns should indicate the subject IDs, rater IDs, and observation scores. See the Example application in Ten Hove et al. (2025, Multivariate Behavioral Research) for an example data set.

- OUTPUT:

Estimated variance components (subjects, raters, residual), Design factors ( $k$ = total number of raters, $\hat{k}$ = harmonic mean number of raters per subject, $Q$ = proportion of non-overlapping raters across subjects; see Ten Hove et al., 2024, Psychological Methods ) and six types of intraclass correlation coefficients (ICCs).

- UNSURE WHICH ICC TO INTERPRET?

Use the Tab `*Flowchart*` .

- ALTERNATIVE DESIGN FACTORS?

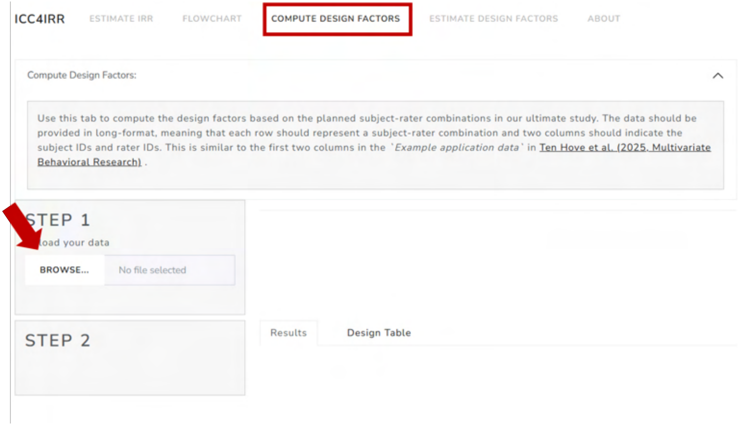
Will your ultimate study use a different observation design than was used in the uploaded data set (e.g., different numbers of raters per subject, or less/more overlapping raters across subjects)? Use the Tab `*Compute Design Factors*` (if you already know the observation design of your ultimate study) or `*Estimate Design Factors*` (if you still need to decide about the observation design of your primary study) to find the relevant values for $\hat{k}$ and $Q$ . Next, use these design factors in the optional Step 3 to estimate the ICC for your ultimate observation study.

STEP 1

Upload your data

© 2025. See 'About' page for citation info.

# ICC4IRR application



Compute Design Factors:

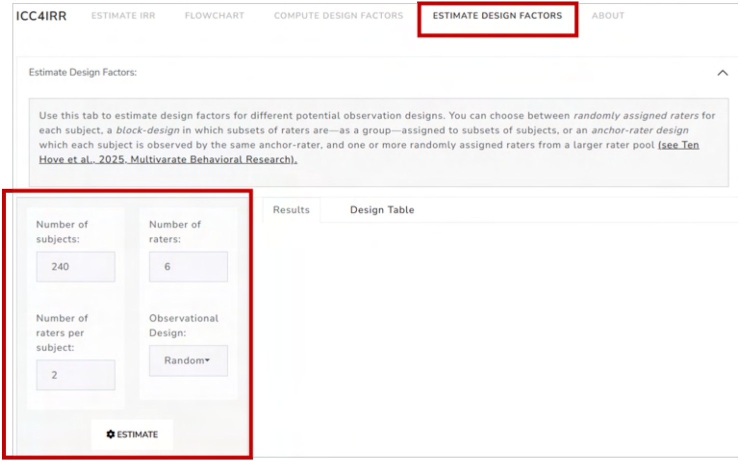
Use this tab to compute the design factors based on the planned subject-rater combinations in our ultimate study. The data should be provided in long-format, meaning that each row should represent a subject-rater combination and two columns should indicate the subject IDs and rater IDs. This is similar to the first two columns in the `*Example application data*` in Ten Hove et al. (2025, Multivariate Behavioral Research) .

STEP 1

Load your data

STEP 2

# ICC4IRR application



Estimate Design Factors:

Use this tab to estimate design factors for different potential observation designs. You can choose between *randomly assigned raters* for each subject, a *block-design* in which subsets of raters are—as a group—assigned to subsets of subjects, or an *anchor-rater design* which each subject is observed by the same anchor-rater, and one or more randomly assigned raters from a larger rater pool (see Ten Hove et al., 2025, Multivarate Behavioral Research).

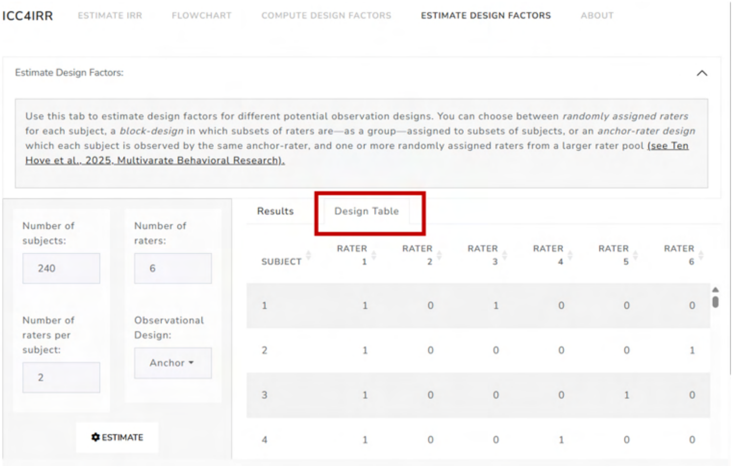
Number of subjects: 240

Number of raters: 6

Number of raters per subject: 2

Observational Design: Random

# ICC4IRR application



Estimate Design Factors:

Use this tab to estimate design factors for different potential observation designs. You can choose between *randomly assigned raters* for each subject, a *block-design* in which subsets of raters are—as a group—assigned to subsets of subjects, or an *anchor-rater design* which each subject is observed by the same anchor-rater, and one or more randomly assigned raters from a larger rater pool (see Ten Hove et al., 2025, Multivarate Behavioral Research).

Number of subjects: 240

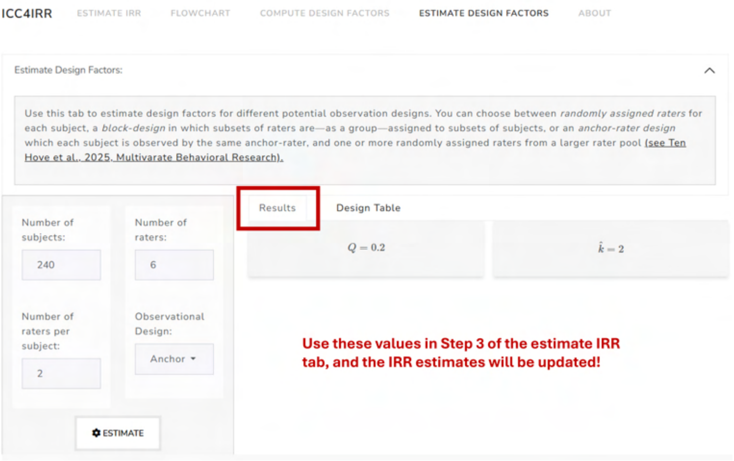
Number of raters: 6

Number of raters per subject: 2

Observational Design: Anchor

| SUBJECT | RATER 1 | RATER 2 | RATER 3 | RATER 4 | RATER 5 | RATER 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 |

# ICC4IRR application



Estimate Design Factors:

Use this tab to estimate design factors for different potential observation designs. You can choose between *randomly assigned raters* for each subject, a *block-design* in which subsets of raters are—as a group—assigned to subsets of subjects, or an *anchor-rater design* which each subject is observed by the same anchor-rater, and one or more randomly assigned raters from a larger rater pool (see Ten Hove et al., 2025, Multivarate Behavioral Research).

Number of subjects: 240

Number of raters: 6

Number of raters per subject: 2

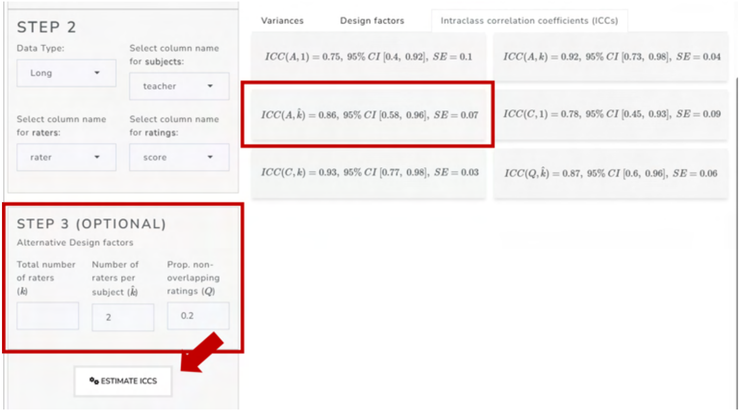
Observational Design: Anchor

$Q = 0.2$

$\hat{k} = 2$

**Use these values in Step 3 of the estimate IRR tab, and the IRR estimates will be updated!**

# ICC4IRR application



STEP 2

Data Type: Long

Select column name for **subjects**: teacher

Select column name for **raters**: rater

Select column name for **ratings**: score

STEP 3 (OPTIONAL)

Alternative Design factors

Total number of raters ($k$):

Number of raters per subject ($\hat{k}$): 2

Prop. non-overlapping ratings ($Q$): 0.2

$ICC(A,1) = 0.75,\ 95\%\ CI\ [0.4,\ 0.92],\ SE = 0.1$

$ICC(A,k) = 0.92,\ 95\%\ CI\ [0.73,\ 0.98],\ SE = 0.04$

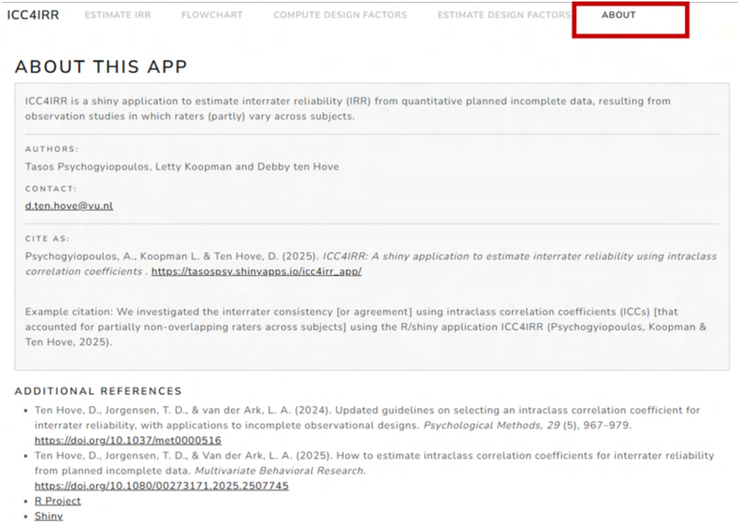
$ICC(A,\hat{k}) = 0.86,\ 95\%\ CI\ [0.58,\ 0.96],\ SE = 0.07$

$ICC(C,1) = 0.78,\ 95\%\ CI\ [0.45,\ 0.93],\ SE = 0.09$

$ICC(C,k) = 0.93,\ 95\%\ CI\ [0.77,\ 0.98],\ SE = 0.03$

$ICC(Q,\hat{k}) = 0.87,\ 95\%\ CI\ [0.6,\ 0.96],\ SE = 0.06$

# ICC4IRR application



## ABOUT THIS APP

ICC4IRR is a shiny application to estimate interrater reliability (IRR) from quantitative planned incomplete data, resulting from observation studies in which raters (partly) vary across subjects.

AUTHORS:

Tasos Psychogyiopoulos, Letty Koopman and Debby ten Hove

CONTACT:

d.ten.hove@vu.nl

CITE AS:

Psychogyiopoulos, A., Koopman L. & Ten Hove, D. (2025). *ICC4IRR: A shiny application to estimate interrater reliability using intraclass correlation coefficients* . https://tasospsy.shinyapps.io/icc4irr_app/

Example citation: We investigated the interrater consistency [or agreement] using intraclass correlation coefficients (ICCs) [that accounted for partially non-overlapping raters across subjects] using the R/shiny application ICC4IRR (Psychogyiopoulos, Koopman & Ten Hove, 2025).

### ADDITIONAL REFERENCES

- Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods, 29* (5), 967–979. https://doi.org/10.1037/met0000516
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2025). How to estimate intraclass correlation coefficients for interrater reliability from planned incomplete data. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2025.2507745
- R Project
- Shiny

# Thanks for your attention!

## Questions or suggestions?

D.ten.Hove@VU.nl

All software code on **GITHUB**:

SCAN ME

ICC4IRR Shiny app:

SCAN ME

# Key References

- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30
- Psychogiopoulos, T., Koopman, L., & Ten Hove, D. (2025). *ICC4IRR: A shiny application to estimate interrater reliability using intraclass correlation coefficients.* https://tasospsy.shinyapps.io/icc4irr_app
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*. https://psycnet.apa.org/doi/10.1037/0033-2909.86.2.420
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods.* https://doi.org/10.1037/met0000516
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2025). How to Estimate Intraclass Correlation Coefficients for Interrater Reliability from Planned Incomplete Data. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2025.2507745