An Illustration of Advanced Intraclass Correlations for Inter-Rater Reliability

Aaron Matthew Simmons & Jeffrey M. Girard
Department of Psychology, University of Kansas

Background and Introduction

- Inter-rater reliability can be defined as the extent to which raters assigned similar scores to objects of measurement (e.g., diagnoses to patients, types to documents, points to athletes, or stars to movies)
- Intraclass correlation coefficients (**ICCs**) are derived from fractions of variance components and have long been used to **quantify inter-rater reliability** for continuous/near-continuous scores
- Recent advances have extended ICCs to accommodate missing data, unbalanced designs, and multilevel structures (ten Hove, Jorgensen, & van der Ark, 2021, 2022)
- We illustrate the use of advanced two-way ICC formulations to answer important questions about the reliability of ratings in real-world settings
- We also showcase novel software (the ***varde*** R package) that we created to allow users to easily implement these tools and techniques
- Finally, we propose and solicit feedback on our plans for new extensions of the ICC framework including the use of generalized linear (mixed) models to accommodate ratings that are not normally distributed (e.g., binary, ordinal, or bounded)

Questions to Answer when Selecting a Two-way ICC Formulation

1 Are the ratings ultimately used for absolute or relative inferences?

Absolute

Ratings are compared to a fixed criterion

e.g., a grade of 10+ is needed to pass

Yields ICCs called "Agreement"

Relative

Object rankings are of primary interest

e.g., the highest 3 will be accepted

Yields ICCs called "Consistency"

Agreement ICCs ≤ Consistency ICCs

2 Are the ratings ultimately used from single raters or the average of many?

Single

Multiple raters may be too expensive

Estimate the reliability of a single rater

Yields ICCs called "Single-Measures"

Average

The mean of multiple raters is more reliable

Estimate the reliability of averaging k raters

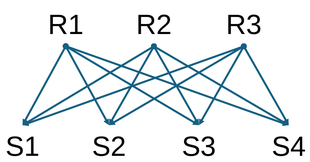
Yields ICCs called "Average-Measures"

Single Measures ICCs ≤ Average-Measures ICCs

3 Was every object rated by every rater, i.e., is the data complete or incomplete?

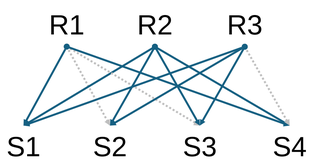
Complete

Fully crossed designs (all possible ratings)



Incomplete

Due to data loss, cost, design, etc.



Raters May be Unbalanced when Incomplete

Name and Error Term for Each Two-way ICC Formulation

Name	Inferences	Measures	Design	Error Term
$\mathrm{ICC}(A,1)$	Absolute	Single	Complete	$\sigma_r^2 + \sigma_{sr}^2$
$\mathrm{ICC}(A,1)$	Absolute	Average	Incomplete	$\sigma_r^2 + \sigma_{sr}^2$
$\mathrm{ICC}(A,k)$	Absolute	Average	Complete	$(\sigma_r^2 + \sigma_{sr}^2)/k$
$\mathrm{ICC}(A,\hat{k})$	Absolute	Average	Incomplete	$(\sigma_r^2 + \sigma_{sr}^2)/\hat{k}$

Name	Inferences	Measures	Design	Error Term
$\mathrm{ICC}(C,1)$	Relative	Single	Complete	σ_{sr}^2
$\mathrm{ICC}(Q,1)$	Relative	Single	Incomplete	$q\sigma_r^2 + \sigma_{sr}^2$
$\mathrm{ICC}(C,k)$	Relative	Average	Complete	σ_{sr}^2/k
$\mathrm{ICC}(Q,\hat{k})$	Relative	Average	Incomplete	$q\sigma_r^2 + \sigma_{sr}^2/\hat{k}$

Introducing the {varde} R package for Variance Decomposition

Introduction and Background

{varde} is a new open-source R package for variance decomposition

It can calculate all two-way ICCs (both frequentist and Bayesian)

Learn more about the software at: https://github.com/affcomlab/varde

Installation and Usage

```
> library(remotes)
> install_github("affcomlab/varde")
> library(varde)
> results <- calc_icc(
    .data = ppa_type1, # example dataset
    subject = "Target", # subject variable
    rater = "Rater", # rater variable
    scores = "Score", # score variable
    k = 12, # number of raters to average
    cores = 4 # optional multicore support
  )
```

Results Summary

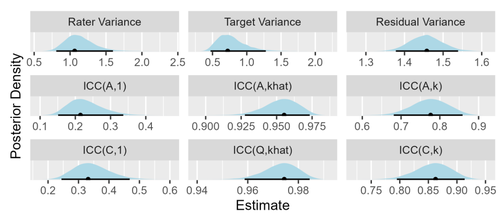
```
> summary(results)

  term         estimate lower upper raters error
  <chr>           <dbl> <dbl> <dbl>  <dbl> <chr>
1 ICC(A,1)        0.220 0.154 0.337      1 Absolute
2 ICC(A,k)        0.782 0.686 0.859     12 Absolute
3 ICC(A,khat)     0.956 0.929 0.973     72 Absolute
4 ICC(C,1)        0.336 0.246 0.470      1 Relative
5 ICC(C,k)        0.862 0.797 0.914     12 Relative
6 ICC(Q,khat)     0.974 0.959 0.985     72 Relative
```

Note that the dataset has 72 raters and is complete, so $\hat{k} = 72$

Results Figure

```
> plot(results)
```



Future Directions

- Add support for multilevel ICCs (ten Hove, Jorgensen, & van der Ark, 2021) to the *varde* package
- Develop new extension to calculate ICCs for non-normal ratings (e.g., using generalized mixed models)
- Enable the calculation of average-measures reliability estimates for discrete scores (on latent scale)

References

- Girard, J.M., & Simmons, A.M. (2024). *varde: R* functions for variance decomposition. R Package version 0.0.1. https://github.com/affcomlab/varde
- ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods.* Advanced online publication.
- ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, 27(4), 650–666.