When intraclass correlation is not suited for measuring test-retest reliability

Gang Chen[1], Daniel S. Pine[2], Melissa A. Brotman[3], Ashley R. Smith[2], Robert W. Cox[1] and Simone P. Haller[2]

1. Scientific and Statistical Computing Core, National Institute of Mental Health, NIH, USA
2. Section on Development and Affective Neuroscience, National Institute of Mental Health, NIH, USA
3. Neuroscience and Novel Therapeutics Unit, Emotion and Development Branch, National Institute of Mental Health, NIH, USA

Correspondence: gangchen@mail.nih.gov

Overview

Test-retest reliability (TRR)

- consistency of an effect across time
- critical criterion for studies of individual differences
- conventional metric: intraclass correlation (ICC)
- poor ICC reported in literature
  - neuroimaging tasks: less than 0.4 (Elliott et al, 2020)
  - behavior data: around 0.5 or below (Hedge et al, 2018)

Main findings regarding ICC based on current investigation

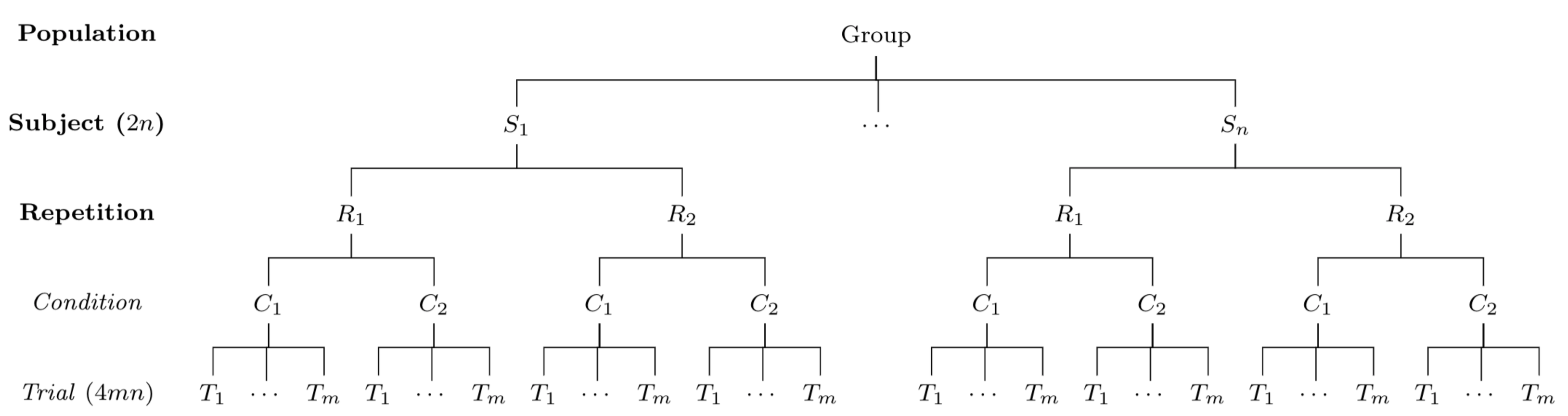
- Conventional ICC is unsuited for test-retest reliability due to its underestimation
  - lower the trial sample size, worse the ICC underestimation
  - higher the cross-trial relative to cross-subject variability, worse the ICC underestimation

Suggestions for test-retest reliability assessments

- construct hierarchical model that explicitly accounts for cross-trial variability
- design an experiment with a large number of trials
- two programs are available in AFNI for test-retest reliability estimation
  - **TRR**: region-level, behavior
  - **3dLMEr**: whole-brain voxel-level

Modeling framework

Typical data hierarchy for test-retest reliability



- Data y_{crst}; subject $s = 1, 2, ..., n$; trial $t = 1, 2, ..., m$; condition $c = 1, 2$; session $r = 1, 2$
- Efect of interest: contrast between two conditions

ICC formulation

- Data aggregation across trials

$$\widehat{y}_{crs\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{crst},\; c = 1, 2; r = 1, 2;\; s = 1, 2, ..., n$$

$$y_{rs} = \widehat{y}_{1rs\cdot} - \widehat{y}_{2rs\cdot}$$

- Conventional model formulation for ICC

$$y_{rs} \sim \mathcal{N}(a_r + \tau_s,\; \sigma_e^2);\; \tau_s \sim \mathcal{N}(0,\; \widetilde{\sigma}_\tau^2);\; r = 1, 2;\; s = 1, 2, ..., n$$

- ICC as variance ratio or correlation between sessions

$$\text{ICC(3,1)} = \frac{\widetilde{\sigma}_\tau^2}{\widetilde{\sigma}_\tau^2 + \sigma_e^2}.$$

Problems with ICC formulation

- trial-level effects: not explicitly accounted for
- data generating mechanism: not accurately characterized
- uncertainty ignored in real practice

Hierarchical model for test-retest reliability (Chen et al., 2021)

$$y_{crst} \sim \mathcal{N}(\mu_{crs},\; \sigma_0^2);\; \mu_{crs} = a_r + b_r I_c + \tau_{rs} + \lambda_{rs} I_c;$$

$$(\tau_{1s},\; \tau_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times1},\; \boldsymbol{R}^{(0)}_{2\times2});\; (\lambda_{1s},\; \lambda_{2s})^T \sim \mathcal{N}(\mathbf{0}_{2\times1},\; \boldsymbol{R}^{(1)}_{2\times2});$$

$$\boldsymbol{R}^{(0)} = \begin{bmatrix} \sigma_{\tau_1}^2 & \rho_0\sigma_{\tau_1}\sigma_{\tau_2} \\ \rho_0\sigma_{\tau_1}\sigma_{\tau_2} & \sigma_{\tau_2}^2 \end{bmatrix};\; \boldsymbol{R}^{(1)} = \begin{bmatrix} \sigma_{\lambda_1}^2 & \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} \\ \rho_1\sigma_{\lambda_1}\sigma_{\lambda_2} & \sigma_{\lambda_2}^2 \end{bmatrix};\; I_c = \begin{cases} \frac{1}{2}, & \text{if } c = 1; \\ -\frac{1}{2}, & \text{if } c = 2. \end{cases}$$

$$c = 1, 2;\; r = 1, 2;\; s = 1, 2, .., n;\; t = 1, 2, .., m.$$

Revelations from the hierarchical model for test-retest reliability

- ρ_0: test-retest reliability for the average between the two conditions
- ρ_1: test-retest reliability for the contrast between the two conditions
- ICC underestimation: cross-trial variability $\frac{2}{m}\sigma_0^2$ unaccounted for in conventional model
- ICC underestimation ratio $\frac{1}{1+\frac{2}{m}R_v^2}$; $R_v = \frac{\sigma_0}{\sigma_\lambda}$: cross-trial relative to cross-subject variability

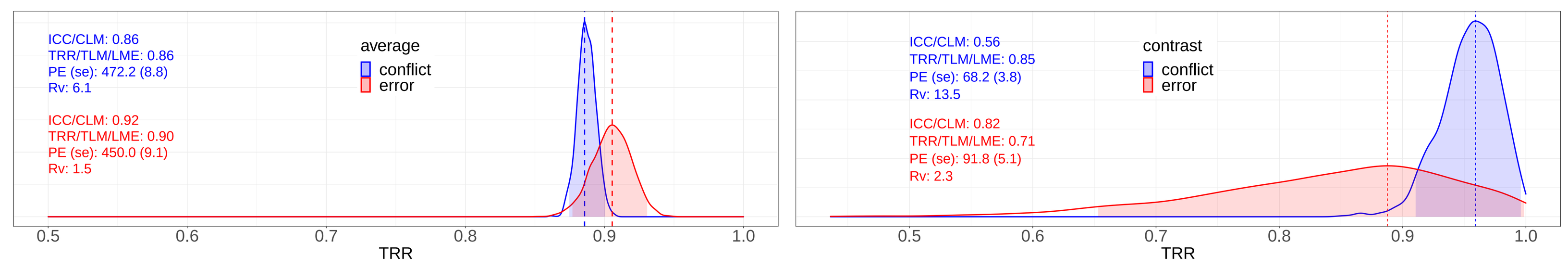
Applications to an experimental dataset

Data structure

- Flanker task: 2 conditions (congruent and incongruent); 2 sessions
- conflict effect for correct responses: $n = 42$ subjects; $m = 350 \pm 36$ incongruent trials and $m = 412 \pm 19$ congruent trials
- error effect for incorrect responses: $n = 27$ subjects; $m = 331 \pm 28$ incongruent correct trials and $m = 90 \pm 27$ incongruent commission error trials
- effects of interest
  - average and contrast of reaction time between congruent and incongruent conditions
  - average and contrast of correct responses between congruent and incongruent conditions
- 12 regions of interest: cognitive control (6); default mode (4); visual (2)

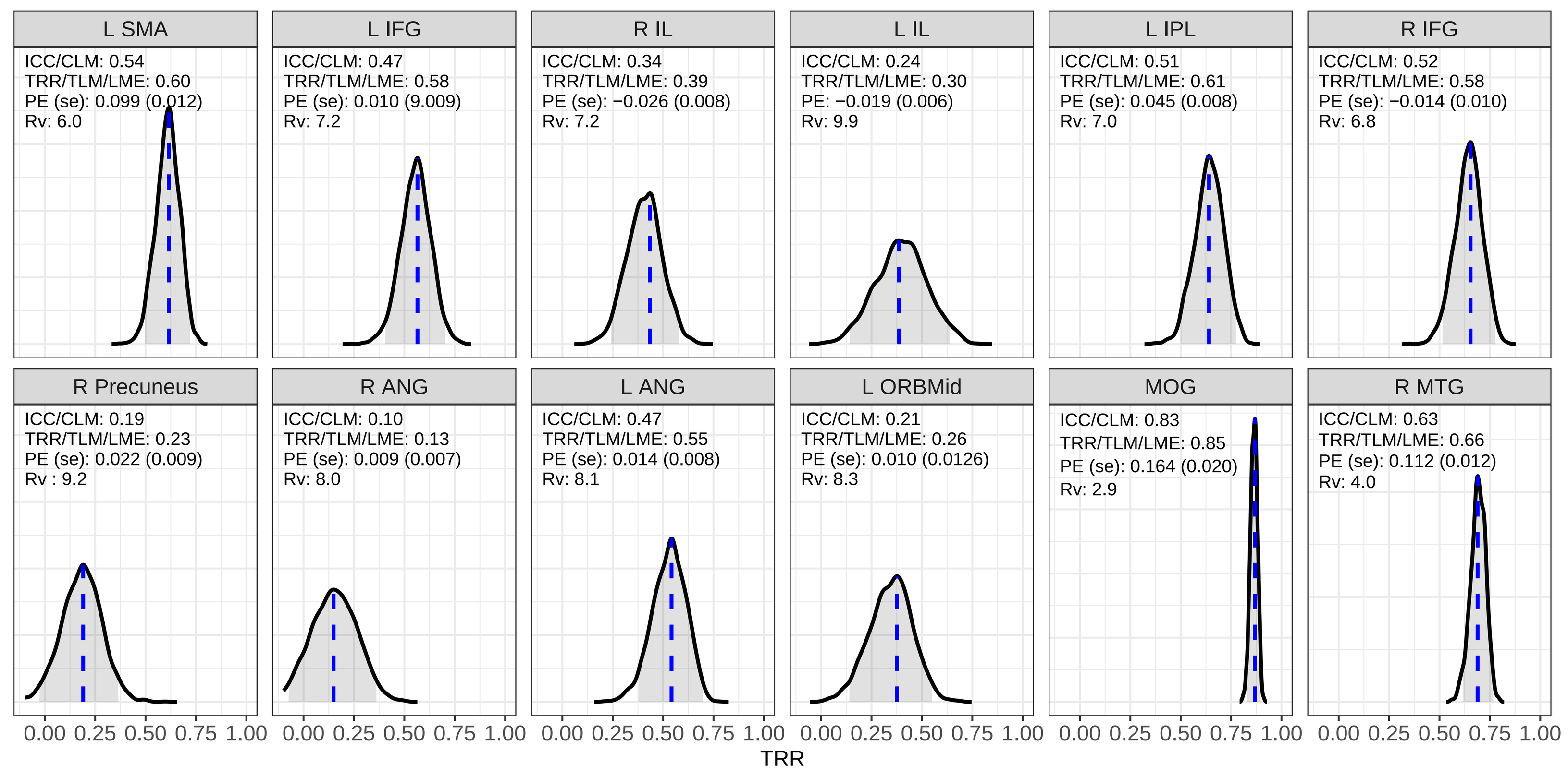
Abbr.	Region	Abbr.	Region	Abbr.	Region
SMA	supplementary motor area	IFG	inferior frontal gyrus	IL	insula lobe
IPL	inferior parietal lobule	PreCG	precentral gyrus	MOG	middle occipital gyrus
MTG	middle temporal gyrus	ANG	angular gyrus	ORBmid	middle orbital gyrus

Test-retest reliability for reaction time



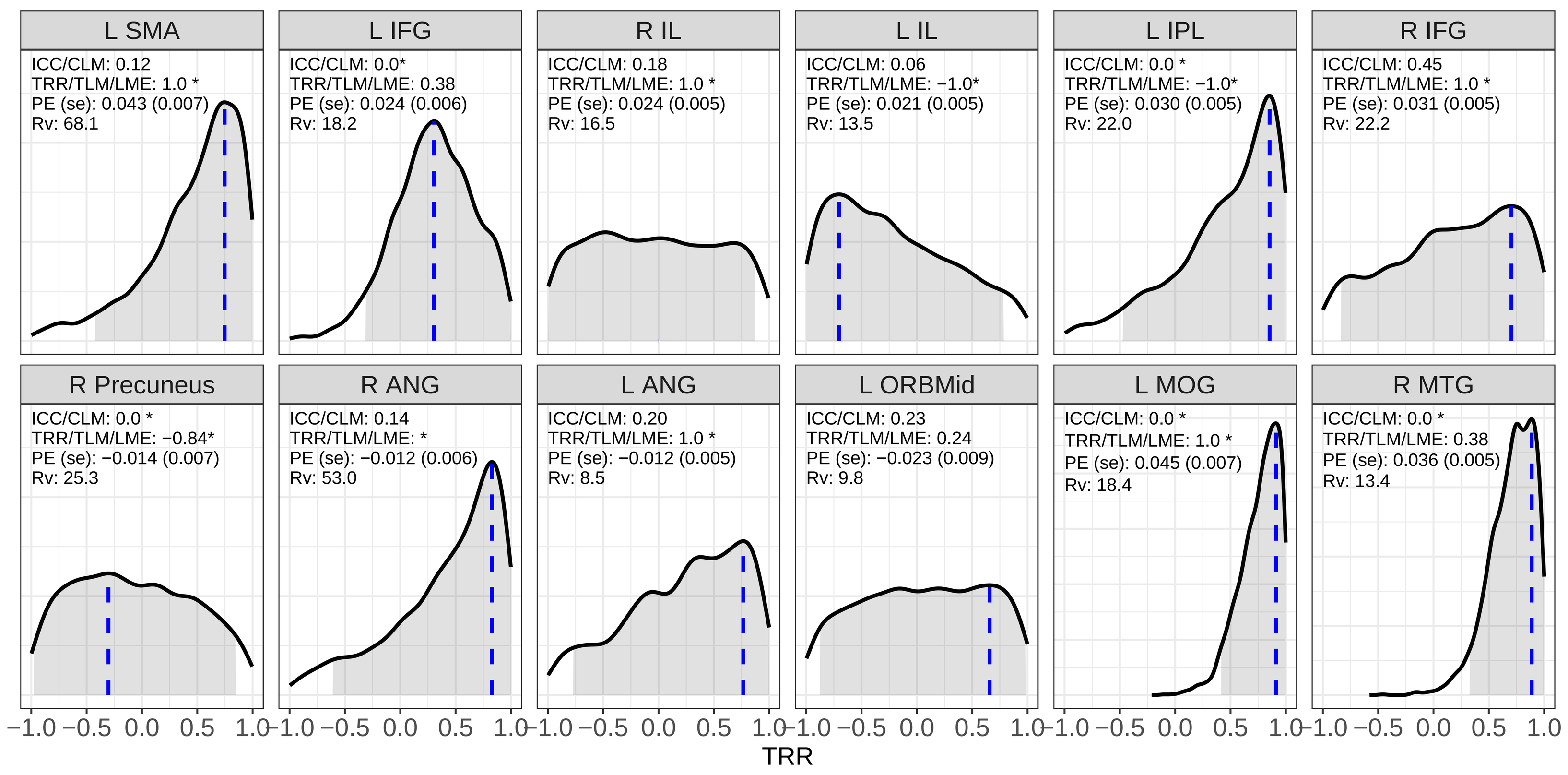
- TRR: high with low uncertainty for average effect and moderate uncertainty for contrast
- ICC underestimation: negligible for average effect, but sizeable for contrast
- Cross-trial variability ratio R_v: roughly same order as cross-subject variability for error effect, but much higher for conflict effect

Test-retest reliability for neuroimaging data: average of correct responses between congruent and incongruent conditions



- TRR: varying across regions with moderate to high precision
- ICC underestimation: negligible to moderate
- Cross-trial variability ratio R_v: moderate to high

Test-retest reliability for neuroimaging data: contrast of correct responses between congruent and incongruent conditions



- TRR: varying across regions with poor to moderate precision
- ICC underestimation: substantial
- Cross-trial variability ratio R_v: very high

Summary

ICC: unsuited for assessing test-retest reliability with data of multiple trials

- ICC tends to underestimates test-retest reliability
- Lower the trial sample size, worse the underestimation
- Larger the cross-trial variability, worse the underestimation
- Converging evidence shows substantially large cross-trial variability
  - reaction time in psychometrics: $3 \leq R_v \leq 11$
  - FMRI: $10 \leq R_v \leq 100$
- Uncertainty information for ICC estimation is usually not reported in literature

Hierarchical modeling platform: more appropriate for test-retest reliability

- Bayesian framework allows for flexibility
- incorporation of uncertainty for effect estimates
- wide range of distribution adaptivity (Gaussian, exGaussian, Student, log-normal, ...)
- availability of estimate precision
- a large number of trials (e.g., hundreds) needed to achieve a high TRR precision

Two programs: TRR and 3dLMEr available in AFNI for TRR estimation

Acknowledgments

The research was supported by the NIMH & NINDS Intramural Research Programs of the NIH.

References

Chen et al, 2021. Beyond the intraclass correlation: A hierarchical modeling approach to test-retest assessment. bioRxiv 2021.01.04.425305.
Elliott et al, 2020. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychological Science.
Hedge et al, 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav Res 50, 1166–1186.