The effects of reliability on statistical power in simple designs.

keith lohse, phd

school of kinesiology, auburn university

Reliability in classical test theory.

- Observed scores are a function of the underlying “true” value plus error. This error can be decomposed into systematic and random sources, but we can generally think of if at measurement error.

$$X_i = T_i + \varepsilon_i$$

- T and ε are random normal variables and the observed score, X, is the resulting sum. This also leads to another important feature, the sum of the variances is also equal to the variance of the sums.

$$var(X_i) = var(T_i) + var(\varepsilon_i)$$

Reliability in classical test theory.

- This also means that the amount of shared variance between the true scores and the observed scores is a function of the proportion of true variance to error variance.

$$r_{tx}^2 = \frac{var(T_i)}{var(T_i) + var(\varepsilon_i)} = \frac{var(T_i)}{var(X_i)}$$

- Thus by keeping the distribution of T fixed as a standard normal variable, N(0,1), we can manipulate the variance of ε to produce the desired r/r^2 between the true scores and the observed scores.

Reliability in classical test theory.

var(T)	var(X) for des corr	var(e) for des corr	shared_var (r^2)	des_corr (r)
1	4	3	0.25	0.5
1	2.777777778	1.777777778	0.36	0.6
1	2.040816327	1.040816327	0.49	0.7
1	1.5625	0.5625	0.64	0.8
1	1.234567901	0.234567901	0.81	0.9

- Based on these values, we can simulate populations of true scores (N=10,000) from which we can generate the observed X variables.
  - **Independent samples t-test:**
    - Control group true scores: $T \sim N(0,1)$
    - Experimental group true scores: $T \sim N(\delta, 1)$
  - **Paired samples t-test:**
    - Pre-test true scores: $T_{pre} \sim N(0,1)$
    - Post-test true scores: $T_{post} = T_{pre} + \delta$
  - **Within-between interaction in 2x2 ANOVA:**
    - Control group: $T_{pre} \sim N(0,1); T_{post} = T_{pre} + 0$
    - Experimental group: $T_{pre} \sim N(0,1);\ T_{post} = T_{pre} + \delta$

Where δ is the true Cohen's d in the population.

Reliability in classical test theory.

var(T)	var(X) for des corr	var(e) for des corr	shared_var (r^2)	des_corr (r)
1	4	3	0.25	0.5
1	2.777777778	1.777777778	0.36	0.6
1	2.040816327	1.040816327	0.49	0.7
1	1.5625	0.5625	0.64	0.8
1	1.234567901	0.234567901	0.81	0.9

- Measurement error is then added, $\varepsilon_r \sim (0, var(e_r))$ and we have five different populations of observed scores, and one set of true scores, from which we can sample.
- For each analysis, we simulate 10,000 experiments and then look at the proportion of statistically significant results.
  - We do this for nine different sample sizes, n = [10, 20, 30, 40, 50, 60, 100, 200, 300], and two different effect sizes, δ = [0.5, 0.8].

Cells contain proportion of statistically significant results out of 10,000 simulated experiments.

Simulated independent t-test results, Cohen's d = 0.5.						
	Measurement error in dependent variable					
n/group	T (no error)	X r^2_{TX}=0.81	X r^2_{TX}=0.64	X r^2_{TX}=0.49	X r^2_{TX}=0.36	X r^2_{TX}=0.25
10	0.18	0.16	0.13	0.11	0.09	0.08
20	0.32	0.27	0.22	0.19	0.15	0.12
30	0.47	0.39	0.31	0.27	0.20	0.16
40	0.58	0.49	0.39	0.34	0.25	0.19
50	0.69	0.58	0.48	0.41	0.31	0.23
60	0.77	0.66	0.56	0.47	0.37	0.27
100	0.94	0.87	0.76	0.69	0.54	0.41
200	0.99	0.99	0.97	0.94	0.83	0.71
300	1.00	0.99	0.99	0.99	0.95	0.87
Simulated independent t-test results, Cohen's d = 0.8.						
	Measurement error in dependent variable					
n/group	T (no error)	X r^2_{TX}=0.81	X r^2_{TX}=0.64	X r^2_{TX}=0.49	X r^2_{TX}=0.36	X r^2_{TX}=0.25
10	0.40	0.33	0.27	0.22	0.19	0.13
20	0.68	0.61	0.50	0.42	0.34	0.23
30	0.86	0.79	0.68	0.60	0.47	0.33
40	0.94	0.89	0.80	0.71	0.59	0.42
50	0.98	0.94	0.88	0.81	0.69	0.52
60	0.99	0.98	0.93	0.87	0.77	0.58
100	1.00	1.00	0.99	0.98	0.94	0.79
200	1.00	1.00	1.00	1.00	1.00	0.98
300	1.00	1.00	1.00	1.00	1.00	1.00

Cells contain proportion of statistically significant results out of 10,000 simulated experiments.

Simulated paired t-test results, Cohen's d = 0.5.						
	Measurement error in dependent variable					
Total N =	$T_{pre/post}$ (no error)	$X_{pre/post}$ r^2_{TX}=0.81	$X_{pre/post}$ r^2_{TX}=0.64	$X_{pre/post}$ r^2_{TX}=0.49	$X_{pre/post}$ r^2_{TX}=0.36	$X_{pre/post}$ r^2_{TX}=0.25
10	1.00	0.53	0.26	0.16	0.11	0.09
20	1.00	0.86	0.49	0.32	0.20	0.14
30	1.00	0.97	0.68	0.46	0.28	0.19
40	1.00	0.99	0.81	0.58	0.36	0.24
50	1.00	1.00	0.88	0.69	0.44	0.29
60	1.00	1.00	0.94	0.77	0.51	0.33
100	1.00	1.00	1.00	0.93	0.74	0.51
200	1.00	1.00	1.00	1.00	0.96	0.82
300	1.00	1.00	1.00	1.00	0.99	0.94
Simulated paired t-test results, Cohen's d = 0.8.						
	Measurement error in dependent variable					
Total N	$T_{pre/post}$ (no error)	$X_{pre/post}$ r_{TX}=0.81	$X_{pre/post}$ r_{TX}=0.64	$X_{pre/post}$ r_{TX}=0.49	$X_{pre/post}$ r_{TX}=0.36	$X_{pre/post}$ r_{TX}=0.25
10	1.00	0.90	0.55	0.35	0.22	0.15
20	1.00	0.99	0.89	0.66	0.43	0.28
30	1.00	0.99	0.98	0.84	0.60	0.41
40	1.00	1.00	1.00	0.93	0.74	0.51
50	1.00	1.00	1.00	0.97	0.83	0.62
60	1.00	1.00	1.00	0.99	0.90	0.69
100	1.00	1.00	1.00	1.00	0.98	0.90
200	1.00	1.00	1.00	1.00	1.00	0.99
300	1.00	1.00	1.00	1.00	1.00	1.00

Cells contain proportion of statistically significant results out of 10,000 simulated experiments.

Simulated interaction results, post-test Cohen's d = 0.5 (no pre-test difference).						
	Measurement error in dependent variable					
n/group =	$T_{pre/post}$ (no error)	$X_{pre/post}$ r^2_{TX}=0.81	$X_{pre/post}$ r^2_{TX}=0.64	$X_{pre/post}$ r^2_{TX}=0.49	$X_{pre/post}$ r^2_{TX}=0.36	$X_{pre/post}$ r^2_{TX}=0.25
10	1.00	0.34	0.16	0.12	0.09	0.08
20	1.00	0.62	0.29	0.19	0.13	0.10
30	1.00	0.79	0.40	0.28	0.17	0.12
40	1.00	0.90	0.51	0.34	0.23	0.15
50	1.00	0.95	0.61	0.42	0.26	0.18
60	1.00	0.98	0.69	0.48	0.32	0.21
100	1.00	1.00	0.89	0.71	0.48	0.32
200	1.00	1.00	0.99	0.94	0.78	0.55
300	1.00	1.00	1.00	0.99	0.91	0.73
Simulated interaction results, post-test Cohen's d = 0.8 (no pre-test difference).						
	Measurement error in dependent variable					
n/group =	$T_{pre/post}$ (no error)	$X_{pre/post}$ r^2_{TX}=0.81	$X_{pre/post}$ r^2_{TX}=0.64	$X_{pre/post}$ r^2_{TX}=0.49	$X_{pre/post}$ r^2_{TX}=0.36	$X_{pre/post}$ r^2_{TX}=0.25
10	1.00	0.70	0.35	0.22	0.14	0.11
20	1.00	0.95	0.62	0.40	0.27	0.18
30	1.00	0.99	0.80	0.57	0.37	0.24
40	1.00	0.99	0.90	0.69	0.47	0.31
50	1.00	1.00	0.95	0.79	0.57	0.38
60	1.00	1.00	0.98	0.86	0.64	0.44
100	1.00	1.00	0.99	0.98	0.85	0.65
200	1.00	1.00	1.00	0.99	0.99	0.92
300	1.00	1.00	1.00	1.00	1.00	0.98

Relationship between the Intraclass Correlation Coefficient (ICC) and r^2_{tx}.

All the different ICCs

- Broadly speaking, the ICC is a measure of reliability, either between raters or between different time-points. There are several different formulae for ICCs, but the most conceptual interpretation is:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Any score in our data (Y_{ij}) can be expressed as the grand mean (μ) plus a deviate for each 'unit' (α_j; in our case units are individuals) and another deviate, assumed to be random error, for each observation in that unit (ε_{ij}; in our case, time-points within individuals).

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

The ICC can then be expressed as variation between units (i.e., between individuals) relative to the total variation (i.e., variation between individuals plus the variation within individuals).

All the different ICCs

- That said, there are several different methods for calculating and presenting ICCs that are most appropriate in different contexts.
  - See Landers (2015) for a review.

- The table below presents how these ICCs are labelled in SPSS and a conceptual definition of each.

	SPSS Label	Formal Notation	Conceptual Definition
	One Way Random	ICC(1,1) ICC(1,k)	How much variance is shared by the set of *k* measures and the true, underlying construct. Assumes that different raters/times are used for different measures (does not disentangle the effects of the rater and ratee).
*	**Two-Way Random Single Measures**	**ICC(2,1)**	How much variance is shared by a single measure and the true, underlying construct. Assumes raters/times are a *sample*.
*	**Two-Way Random Average Measures**	**ICC(2,k)**	How much variance is shared by the set of *k* measures and the true, underlying construct. Assumes raters/times are a *sample*.
	Two-Way Mixed Single Measures	ICC(3,1)	How much variance is shared by a single measure and the true, underlying construct. Assumes raters/times are a *population*.
	Two-Way Mixed Average Measures	ICC(3,k)	How much variance is shared by the set of *k* measures and the true, underlying construct. Assumes raters/times are a *population*.

Two-Way Random ICCs (2,1) and (2,k)

- Currently, we report the average measures reliability, **ICC (2,k)**, where k = 3 for our three time points. This tells us how much variation in the underlying construct is captured by our three different time-points as a set.
  - This is good because it is the commonly reported ICC. It also means that across three different time points we can actually get a reliable measure of the underlying construct, e.g. if ICC(2,k) ≥ 0.8.
  - However, we cannot use ICC(2,k) for our power analyses, because ICC(2,k) changes as function of the number of time points.
    - I.e., r_{tx}^2 might = 0.25 in the population. If we take only two measurements, the ICC(2,k) might be poor… but if we take ten measurements, the ICC(2,k) might be really good!

Two-Way Random ICCs (2,1) and (2,k)

- As such, I think we need to add **ICC(2,1)** to our results, because it will allow us to translate between our TMS data and r_{tx}^2 in our power analyses.
- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
  - I.e., ICC(2,k) is always going to be higher than ICC(2,1) because averaging across multiple measurements can help us cancel out noise.

llr_latency1 (good reliability)

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.917[a]	.864	.953	34.175	38	76	.000
Average Measures	.971	.950	.984	34.175	38	76	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

cortical_relay_time (poor reliability)

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.338[a]	.123	.559	2.534	32	64	.001
Average Measures	.605	.297	.792	2.534	32	64	.001

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
- Computationally, ICC(2,1) is an estimate of the long-run average of the off-diagonal in a correlation matrix.
  - E.g., in simulated data where $r_{tx} = 0.50$ and thus $\boldsymbol{r_{tx}^2 = 0.25}$ in the population, if we are taking **2** measurements:

Inter-Item Correlation Matrix

	X_5a	X_5b
X_5a	1.000	.225
X_5b	.225	1.000

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.225[a]	.206	.243	1.580	9999	9999	.000
Average Measures	.367	.342	.392	1.580	9999	9999	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
- Computationally, ICC(2,1) is an estimate of the long-run average of the off-diagonal in a correlation matrix.
  - E.g., in simulated data where $r_{tx} = 0.50$ and thus $\boldsymbol{r_{tx}^2 = 0.25}$ in the population, if we are taking **3** measurements:

Inter-Item Correlation Matrix

	X_5a	X_5b	X_5c
X_5a	1.000	.225	.248
X_5b	.225	1.000	.249
X_5c	.248	.249	1.000

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.241[a]	.228	.253	1.951	9999	19998	.000
Average Measures	.487	.470	.505	1.951	9999	19998	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
- Computationally, ICC(2,1) is an estimate of the long-run average of the off-diagonal in a correlation matrix.
  - E.g., in simulated data where $r_{tx} = 0.50$ and thus $\boldsymbol{r_{tx}^2 = 0.25}$ in the population, if we are taking **5** measurements:

Inter-Item Correlation Matrix

	X_5a	X_5b	X_5c	X_5d	X_5e
X_5a	1.000	.255	.240	.257	.256
X_5b	.255	1.000	.236	.260	.251
X_5c	.240	.236	1.000	.257	.228
X_5d	.257	.260	.257	1.000	.250
X_5e	.256	.251	.228	.250	1.000

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.249[a]	.240	.258	2.658	9998	39992	.000
Average Measures	.624	.612	.635	2.658	9998	39992	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
- Computationally, ICC(2,1) is an estimate of the long-run average of the off-diagonal in a correlation matrix.
  - Or, if $r_{tx} = 0.70$ and thus $\boldsymbol{r_{tx}^2 = 0.49}$ in the population, and we are taking **3** measurements:

Inter-Item Correlation Matrix

	X_7a	X_7b	X_7c
X_7a	1.000	.482	.486
X_7b	.482	1.000	.491
X_7c	.486	.491	1.000

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.486[a]	.475	.498	3.840	9999	19998	.000
Average Measures	.740	.731	.748	3.840	9999	19998	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Conceptually, ICC(2,1) is the average variance shared by a ***single measurement*** and the underlying construct.
- Computationally, ICC(2,1) is an estimate of the long-run average of the off-diagonal in a correlation matrix.
  - Or, if $r_{tx} = 0.90$ and thus $\boldsymbol{r_{tx}^2 = 0.81}$ in the population, and we are taking **5** measurements:

Inter-Item Correlation Matrix

	X_9a	X_9b	X_9c	X_9d	X_9e
X_9a	1.000	.808	.813	.809	.813
X_9b	.808	1.000	.805	.808	.808
X_9c	.813	.805	1.000	.808	.807
X_9d	.809	.808	.808	1.000	.810
X_9e	.813	.808	.807	.810	1.000

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.809[a]	.804	.814	22.163	9998	39992	.000
Average Measures	.955	.953	.956	22.163	9998	39992	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

ICC(2,1) is an estimate of r_{tx}^2

- Thus, while the average measures reliability, **ICC(2,k)**, is good to report because it is a common measure and tells us how much of the true construct is captured by a set of measurements, we cannot use it for our power-analyses.
  - Because ICC(2,k) changes as a function of k.

- As such, I think we should add the single measures reliability, **ICC(2,1)**, to our results, because we can then connect our data to the power analyses.
  - Because ICC(2,1) is an estimate of r_{tx}^2.
  - *At sufficiently large n and k, these two values will be equal.*

  - We need to add ICC(2,1) to the results and change the headings in our tables from r_{tx} to r_{tx}^2.

References

1. Landers, R. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower*, 2, e143518.81744.
2. Shrout, P., & Fleiss (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
3. Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284-290.
4. de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., & Knol, D. (2011). *Measurement in Medicine*. Cambridge, UK: Cambridge University Press