Reliability & Agreement

DeShon - 2006

Internal Consistency Reliability

- Parallel forms reliability
- Split-Half reliability
- Cronbach's alpha – Tau equivalent
- Spearman-Brown Prophesy formula
  - Longer is more reliable

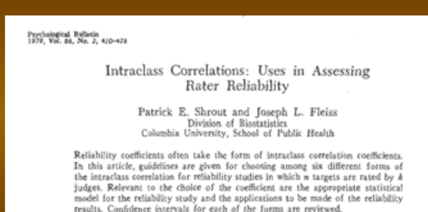
Test-Retest Reliability

- Correlation between the same test administered at two time points
  - Assumes stability of construct
    - Need 3 or more time points to separate error from instability (Kenny & Zarutta, 1996)
  - Assumes no learning, practice, or fatigue effects (tabula rasa)
- Probably the most important form of reliability for psychological inference

Interrater Reliability

- Could be estimated as correlation between two raters or alpha for 2 or more raters
- Typically estimated using intra-class correlation using ANOVA
  - Shrout & Fleiss (1979); McGraw & Wong (1996)

Interrater Reliability



Psychological Bulletin
1979, Vol. 86, No. 2, 420–428

Intraclass Correlations: 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 chooting 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 Correlations

- What is a class of variables?
  - Variables that share a metric and variance
- Height and Weight are different classes of variables.
- There is only 1 Interclass correlation coefficient – Pearson's r.
- When interested in the relationship between variables of a common class, use an Intraclass Correlation Coefficient.

Intraclass Correlations

- An ICC estimates the reliability ratio directly
  - Recall that…

$$r_{xx} = \frac{\sigma_t^2}{\sigma_O^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

- An ICC is estimated as the ratio of variances:

$$ICC = \frac{Var(subjects)}{Var(subjects) + Var(error)}$$

Intraclass Correlations

- The variance estimates used to compute this ratio are typically computed using ANOVA
  - Person x Rater design
  - In reliability theory, classes are persons
    - between person variance
  - The variance within persons due to rater differences is the error

Intraclass Correlations

- Example…depression ratings

Persons	Rater1	Rater2	Rater3	Rater4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

Intraclass Correlations

- 3 sources of variance in the design:
  - persons, raters, & residual error
- No replications so the Rater x Ratee interaction is confounded with the error
- ANOVA results…

Source	df	MS
Between Persons	5	11.24
Within Persons	18	6.26
Between Raters	3	32.49
Residual Error	15	1.02

Intraclass Correlations

- Based on this rating design, Shrout & Fleiss defined three ICCs
  - ICC(1,k) – Random set of people, random set of raters, nested design, rater for each person is selected at random
  - ICC(2,k) – Random set of people, random set of raters, crossed design
  - ICC(3,k) - Random set of people, FIXED set of raters, crossed design

ICC(1,k)

- A set of raters provide ratings on a different sets of persons. No two raters provides ratings for the same person
- In this case, persons are nested within raters.
- Can't separate the rater variance from the error variance
- k refers to the number of judges that will actually be used to get the ratings in the decision making context

ICC(1,k)

$$ICC(1,k) = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_w^2}{k}}$$

- Agreement for the average of k ratings
- We'll worry about estimating these "components of variance" later

ICC(2,k)

$$ICC(2,k) = \frac{\sigma_p^2}{\sigma_p^2 + \frac{(\sigma_r^2 + \sigma_e^2)}{k}}$$

- Because raters are crossed with ratees you can get a separate rater main effect.
- Agreement for the average ratings across a set of random raters

ICC(3,k)

$$ICC(3,k) = \frac{\sigma_p^2}{\sigma_p^2 + \frac{(\sigma_e^2)}{k}}$$

- Raters are "fixed" so you get to drop their variance from the denominator
- Consistency/reliability of the average rating across a set of fixed raters

Shrout & Fleiss (1979)

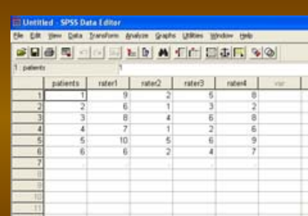
ICC	Estimate
ICC(1,1)	0.17
ICC(2,1)	0.29
ICC(3,1)	0.71
ICC(1,4)	0.44
ICC(2,4)	0.62
ICC(3,4)	0.91

ICCs in SPSS

For SPSS, you must choose:
(1) An ANOVA Model
(2) A Type of ICC

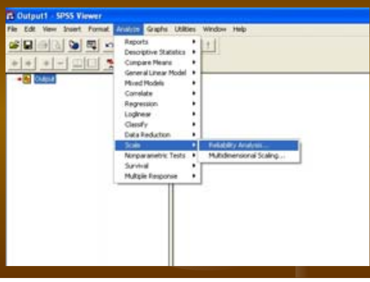
ANOVA Model		
One way Random Effects	ICC(1,1)	
TYPE:	Consistency	Absolute Agreement
Two way Random Effects		**ICC(2,1) "ICC(AGREEMENT)"**
Two way Mixed Model : Raters Fixed Patients Random	ICC(3,1) "ICC(CONSISTENCY)"	

ICCs in SPSS



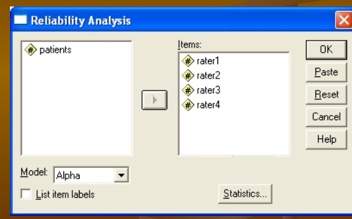
	patients	rater1	rater2	rater3	rater4	var
1	1	9	2	5	8	
2	2	6	1	3	2	
3	3	8	4	6	8	
4	4	7	1	2	6	
5	5	10	5	6	9	
6	6	6	2	4	7	

ICCs in SPSS



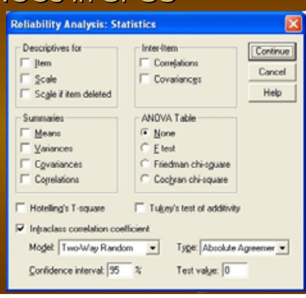
ICCs in SPSS

- Select raters…



ICCs in SPSS

- Choose Analysis under the statistics tab



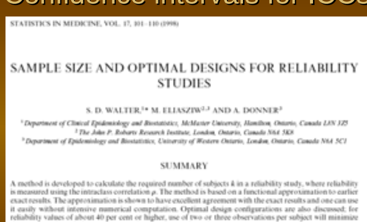
ICCs in SPSS

- Output…

```
              R E L I A B I L I T Y   A N A L Y S I S
                 Intraclass Correlation Coefficient
```

- Two-way Random Effect Model (Absolute Agreement Definition):
- People and Measure Effect Random
- Single Measure Intraclass Correlation = .2898*
- 95.00% C.I.: Lower = .0188 Upper = .7611
- F = 11.02 DF = (5,15.0) Sig. = .0001 (Test Value = .00)
- Average Measure Intraclass Correlation = .6201
- 95.00% C.I.: Lower = .0394 Upper = .9286
- F = 11.0272 DF = (5,15.0) Sig. = .0001 (Test Value = .00)
- Reliability Coefficients
- N of Cases = 6.0 N of Items = 4

Confidence intervals for ICCs


STATISTICS IN MEDICINE, VOL. 17, 101–110 (1998)

SAMPLE SIZE AND OPTIMAL DESIGNS FOR RELIABILITY STUDIES

S. D. WALTER,[1]* M. ELIASZIW[2,3] AND A. DONNER[3]

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5
[2] The John P. Robarts Research Institute, London, Ontario, Canada N6A 5K8
[3] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1

SUMMARY

A method is developed to calculate the required number of subjects k in a reliability study, where reliability is measured using the intraclass correlation ρ. The method is based on a functional approximation to earlier exact results. The approximation is shown to have excellent agreement with the exact results and one can use it easily without intensive numerical computation. Optimal design configurations are also discussed; for reliability values of about 40 per cent or higher, use of two or three observations per subject will minimize the total number of observations required. © 1998 John Wiley & Sons, Ltd.

- For your reference…

Standard Error of Measurement

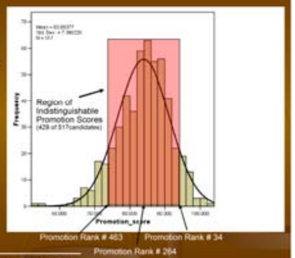
- Estimate of the average distance of observed test scores from an individual's true score.

$$SEM = \sigma_{test} \sqrt{1 - r_{xx}}$$

$$CI = X \pm Z \; SEM$$

Standard Error of the Difference

- Region of indistinguishable true scores



$$SED = \sqrt{SEM_1 + SEM_2}$$

Agreement vs. Reliability

- Reliability/correlation is based on covariance and not the actual value of the two variables
- If one rater is more lenient than another but they rank the candidates the same, then the reliability will be very high
- Agreement requires absolute consistency.

Agreement vs. Reliability

- Interrater Reliability
  - “Degree to which the ratings of different judges are proportional when expressed as deviations from their means” (Tinsley & Weiss, 1975, p. 359)
  - Used when interest is in the relative ordering of the ratings
- Interrater Agreement
  - “Extent to which the different judges tend to make exactly the same judgments about the rated subject” (T&W, p. 359)
  - Used when the absolute value of the ratings matters

Agreement Indices

- Percent agreement
  - What percent of the total ratings are exactly the same?
- Cohen's Kappa
  - Percent agreement corrected for the probability of chance agreement
- r_{wg} – agreement when rating a single stimulus (e.g., a supervisor, community, or clinician).

Kappa

- Typically used to assess interrater agreement
- Designed for categorical judgments (finishing places, disease states)
- Corrects for chance agreements due to limited number of rating scales
  - P_A = Proportion Agreement
  - P_C = expected agreement by chance
- 0 – 1; usually a bit lower than reliability

$$\kappa = \frac{p_A - p_C}{1 - p_C}$$

Kappa Example

		Rater 2 Y	Rater 2 N	
Rater 1	Y	300	20	320
	N	10	70	80
		310	90	400

$$p_A = (300 + 70)/400 = .925$$

Kappa Example

- Expected by chance… $f_e = (M_1 * M_2)/400$

$$p_C = (248+18)/400 = 0.665$$

$$\kappa = \frac{.925+.665}{(1-.665)} = 0.776$$

		Rater 2 Y	Rater 2 N	
Rater 1	Y	248	72	320
	N	62	18	80
		310	90	400

31

Kappa Standards

- Kappa > .8 = good agreement
- .67 <kappa <.8 – "tentative conclusions"
  - Carletta '96
- As with everything…it depends
- For more than 2 raters…
  - Average pairwise kappas

32

Kappa Problems

- Affected by marginals
  - 2 examples with 90% Agreement
  - Ex 1: Kappa = .44
  - Ex 2: Kappa = .80
- Highest kappa with equal amount of yes and no

	Yes	No	
Yes	.85	.05	.90
No	.05	.05	.10
	.90	.10	1.0

	Yes	No	
Yes	.45	.05	.50
No	.05	.45	.50
	.50	.50	1.0

33

Kappa Problems

- Departures from symmetry in the contingency tables (i.e., prevalence and bias) affect the magnitude of kappa.
  - Unbalanced agreement reduces kappa
  - Unbalanced disagreement increases kappa.

34

r_{wg}

- Based on Finn's (1970) index of agreement
- Rwj is used to assess agreement when multple raters rate a single stimulus
- When there is no variation in the stimuli you can't examine the agreement of ratings over different stimuli

35

r_{wg}

- Could use the standard deviation of the ratings
  - Like percent agreement…does account for chance
- r_{wg} references the observed standard deviation in ratings to the expected standard deviation if the ratings are random

36

r_{wg}

- Compares observed variance in ratings to the variance in ratings if ratings were random

$$r_{wg} = 1 - \left(\frac{S_r^2}{\sigma_{EU}^2} \right); \quad where \ \ \sigma_{EU}^2 = (A^2 - 1)/12$$
$$A \ \ is \, the \, No. \, of \ \ scale \ points$$

- Standard assumption is a uniform distribution over the ratings scale range
- .80 - .85 is a reasonable standard