Part 5

Lecture 2 Reliability

Who we are…

Pascal Tyrrell, PhD *Associate Professor*

Department of Medical Imaging , Faculty of Medicine
Department of Statistical Sciences , Faculty of Arts and Science

Paul Corey, PhD *Professor Emeritus*

Biostatistics Program, Dalla Lana Faculty of Public Health
Institute of Medical Science, Faculty of Medicine
Department of Statistical Sciences, Faculty of Arts and Science

RELIABILITY

- Reliability refers to the consistency of a test or measurement.
  - Reliability studies
    - Test-retest reliability
      - Equipment and/or procedures
    - Intra- or inter-rater reliability
      - Assessing the reliability of individual raters or a group of raters

Terminology

❑ Reliability

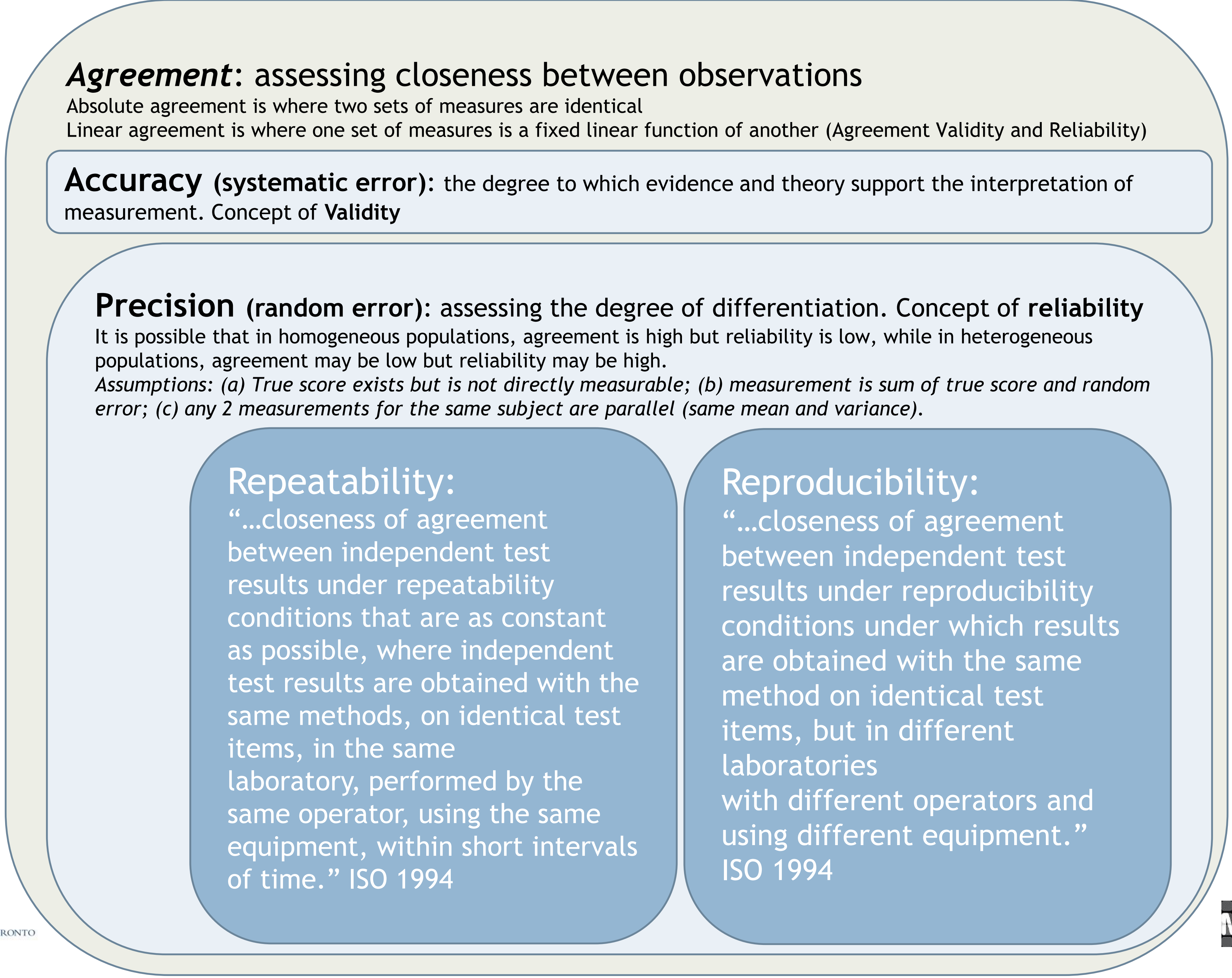
❑ Consistency

❑ Precision

❑ Repeatability

❑ Agreement

❑ “Reliability” and “agreement” are not synonymous!



Agreement: assessing closeness between observations

Absolute agreement is where two sets of measures are identical
Linear agreement is where one set of measures is a fixed linear function of another (Agreement Validity and Reliability)

Accuracy (systematic error): the degree to which evidence and theory support the interpretation of measurement. Concept of **Validity**

Precision (random error): assessing the degree of differentiation. Concept of **reliability**

It is possible that in homogeneous populations, agreement is high but reliability is low, while in heterogeneous populations, agreement may be low but reliability may be high.

Assumptions: (a) True score exists but is not directly measurable; (b) measurement is sum of true score and random error; (c) any 2 measurements for the same subject are parallel (same mean and variance).

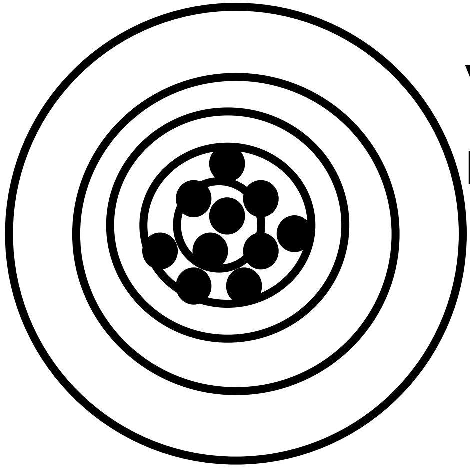
Repeatability:

"…closeness of agreement between independent test results under repeatability conditions that are as constant as possible, where independent test results are obtained with the same methods, on identical test items, in the same laboratory, performed by the same operator, using the same equipment, within short intervals of time." ISO 1994

Reproducibility:

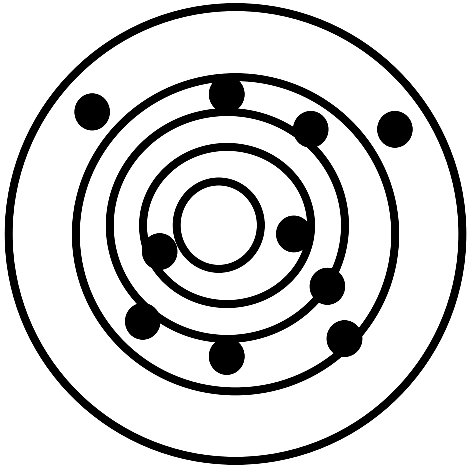
"…closeness of agreement between independent test results under reproducibility conditions under which results are obtained with the same method on identical test items, but in different laboratories
with different operators and using different equipment."
ISO 1994

Validity and Reliability



Valid Yes

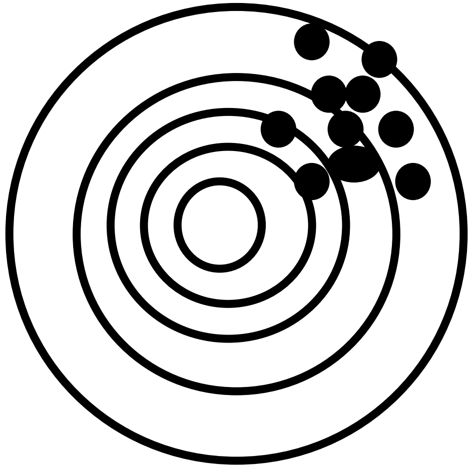
Reliable Yes



Valid No

Reliable No

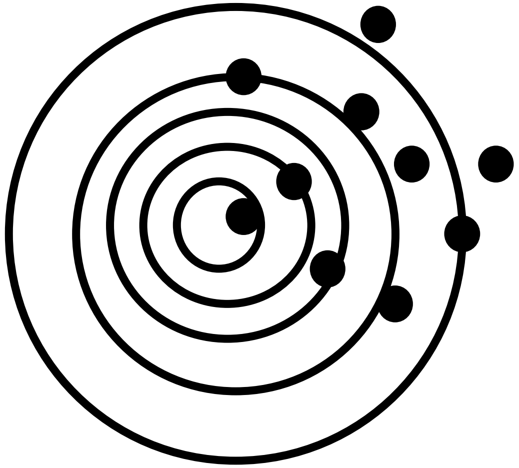
Unbiased



Valid No

Reliable Yes

Biased



Valid No

Reliable No

Biased

Validity

❑ **Success in measuring what you set out to measure**

❑ **Being trained by a Gold Standard trainer ensures validity by**

  ❑ Training on what is proposed to be measured

  ❑ Confirming that everyone is measuring the same thing

Reliability

- ❑ The extent to which the clinical examination yields the same result on repeated inspection.
  - ❑ Inter-examiner reliability:
  - ❑ reproducibility between examiners
  - ❑ Intra-examiner reliability:
  - ❑ reproducibility within examiners

Quantification of Reliability

- In terms of “consistency” of measurements:
  - Relative consistency
    - The consistency of the position or rank of individuals in the group relative to others.
    - Quantified by the “intraclass correlation coefficient” (ICC)
  - Absolute consistency
    - An indication of the “precision” of a score
      - Allows for constructing confidence intervals about a score
    - Quantified by the “standard error of measurement” (SEM) or variations thereof
      - Minimum difference, standard error of prediction (SEP), etc.

Other Procedures Used to Quantify Reliability

- Pearson Product Moment correlation (Pearson r)
  - Cannot detect systematic error

- The coefficient of variation
  - Standard deviation ÷ mean

- Limits of agreement (Bland-Altman plots)
  - Bland-Altman plots compare two measurement techniques on the same variable

Reliability Theory

- ❑ Each observed score is composed of two parts:
  - ❑ True score: the mean of an infinite number of scores from a subject
  - ❑ Error: true score – observed score = error
    - ❑Sources of error:
      - ❑Biological variability, instrumentation, error by the subject, or by the tester, etc.

- ❑ Similarly, for a group of scores, the total variance (σ^2_T) in the data has two components:
  - ❑ True score variance (σ^2_t) and Error variance (σ^2_e)

Medical Imaging
UNIVERSITY OF TORONTO

MiDATA

Reliability Theory

❑ Therefore:

$$\sigma^2{}_{\mathrm{T}} = \sigma^2{}_{\mathrm{t}} + \sigma^2{}_{\mathrm{e}}$$

❑ If we make a ratio of the true score variance ($\sigma^2{}_t$) to the total variance ($\sigma^2{}_T$) we have a reliability coefficient defined as:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

Reliability Theory

- ❑ The closer to 1.0, the higher the reliability
- ❑ Problem…
  - ❑ We don’t actually know the “true score” for each subject; therefore, we don’t know the “true score variability.”
  - ❑ We use an index for true score variability (σ^2_t) based on between-subjects variability; therefore, the formal definition of reliability becomes…

$$R = \frac{\text{Between subjects variabili ty}}{\text{Between subjects variabili ty} + \text{Error}}$$

Variance Estimates

- ❑ Variance estimates are derived from the single-factor, within-subjects ANOVA model
  - ❑ Appropriate mean square values (MS) are recorded from the ANOVA table

  - ❑ NOTE: These will be the values we use to calculate the ICCs

Intraclass Correlation Coefficients

- ICC is a relative measure
  - Ratio of variances from ANOVA
  - Unitless with 1 = perfect reliability; 0 = no reliability
- The relative nature of the ICC and the magnitude of the ICC is dependent on the between-subjects variability
  - ↑ between-subjects variability = ↑ ICC
  - ↓ between-subjects variability = ↓ ICC
    - Therefore, ICCs are context-specific
- “There is literally no such thing as the reliability of a test, unqualified; the coefficient has meaning only when applied to specific populations” Streiner & Norman (1995).

Error

- Two types of error
  - Systematic error
  - Random error

    - Where:
  - *systematic error + random error = total error*

Calculations of Reliability

- ❑ We are interested in calculating the ICC
  - ❑ First step:
    - ❑ Conduct a single-factor, within-subjects (repeated measures) ANOVA
      - ❑ This is an inferential test for systematic error
      - ❑ All subsequent equations are derived from the ANOVA table

ANOVA Table

- ❑ 3 sources of variability
  - ❑ Subjects (MS_B) or (MS_S)
    - ❑ Between-subjects variability (for calculating the ICC)
  - ❑ Trials (MS_T)
    - ❑ Systematic error (for calculating the ICC)
  - ❑ Error (MS_E)
    - ❑ Random error (for calculating the ICC)

- ❑ 2 factors
  - ❑ Trials
    - ❑ Differences between trials
  - ❑ Subjects
    - ❑ Differences between subjects

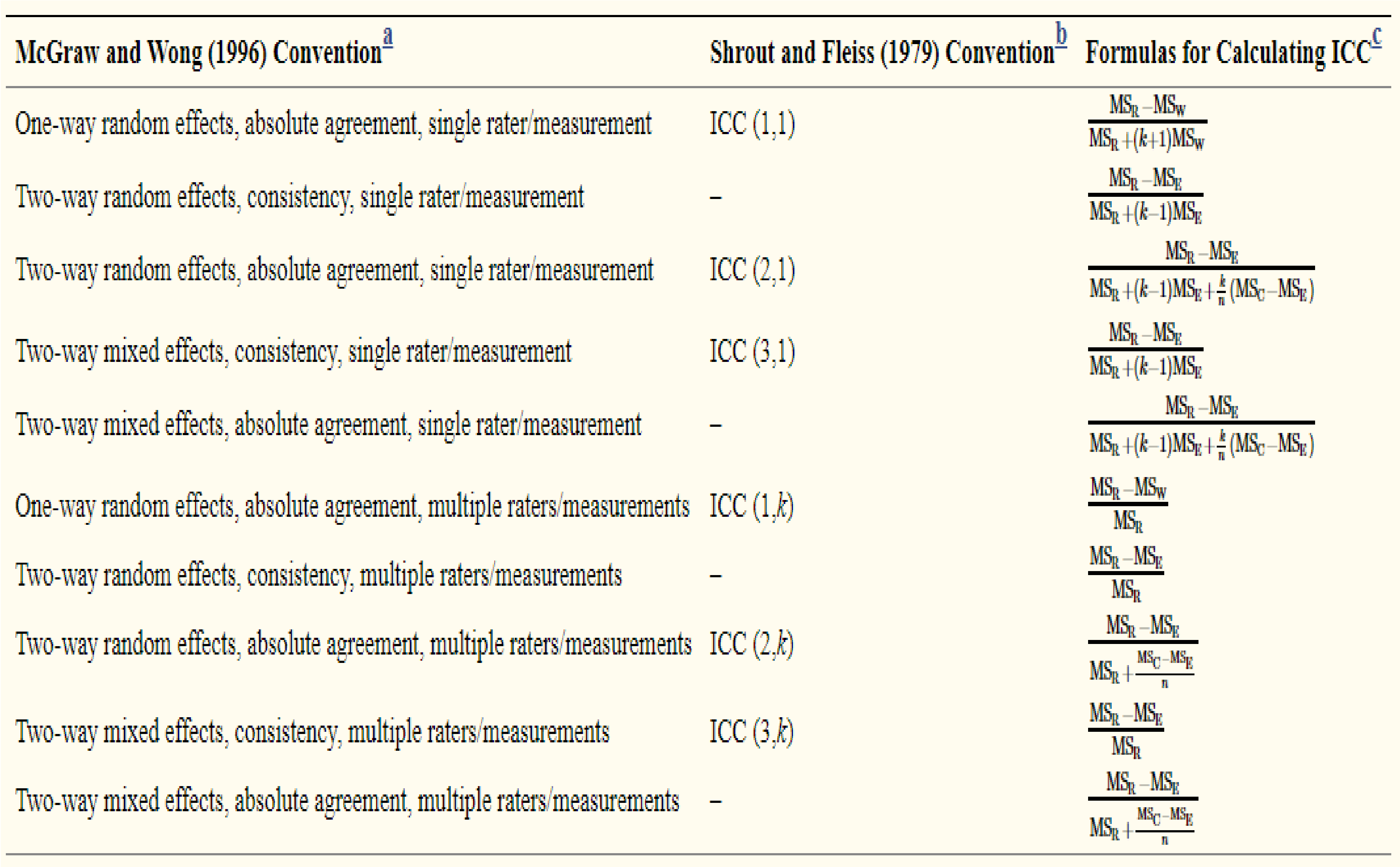
- ❑ Interaction term = *trials x subjects*

ANOVA Table

- ❑ 2 reasons for noting the three different sources of variability
  - ❑ As we will see, there are 6 different ICC models
    - ❑ Two are “one-way models” and four are “two-way models”
      - ❑ One-way models lump together the “trial” and “error” variability
      - ❑ Two-way models keep them separate
  - ❑ Between-subjects ANOVAs are different than within-subjects ANOVAs
    - ❑ The variability due to subjects is not accounted for in the within-subjects ANOVA (due to the repeated testing of the same subject, we assume the same between-subjects variability)

ICC Models

- ❑ Shrout & Fleiss (1979) have developed 6 forms of the ICC:
  - ❑ There are 3 general models:
    - ❑ Models 1, 2, and 3
    - ❑ Each can be calculated two different ways
      - ❑ If the individual scores are actually “single” scores from each subject for each trial, the ICC model is given a second designation of “1”
      - ❑ If the scores in the analysis represent the average of “k” scores from each subject, the ICC is given a second designation of “k”



McGraw and Wong (1996) Convention[a]	Shrout and Fleiss (1979) Convention[b]	Formulas for Calculating ICC[c]
One-way random effects, absolute agreement, single rater/measurement	ICC (1,1)	$\frac{MS_R - MS_W}{MS_R + (k+1)MS_W}$
Two-way random effects, consistency, single rater/measurement	–	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$
Two-way random effects, absolute agreement, single rater/measurement	ICC (2,1)	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$
Two-way mixed effects, consistency, single rater/measurement	ICC (3,1)	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$
Two-way mixed effects, absolute agreement, single rater/measurement	–	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$
One-way random effects, absolute agreement, multiple raters/measurements	ICC (1,*k*)	$\frac{MS_R - MS_W}{MS_R}$
Two-way random effects, consistency, multiple raters/measurements	–	$\frac{MS_R - MS_E}{MS_R}$
Two-way random effects, absolute agreement, multiple raters/measurements	ICC (2,*k*)	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$
Two-way mixed effects, consistency, multiple raters/measurements	ICC (3,*k*)	$\frac{MS_R - MS_E}{MS_R}$
Two-way mixed effects, absolute agreement, multiple raters/measurements	–	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$

ICC Models

- ❑ Usually presented in the context of determining rater reliability
  - ❑ Model 1 (1,1 & 1,k)
    - ❑ Each subject is assumed to be assessed by a different set of raters than other subjects
      - ❑ Random effect of raters
  - ❑ Model 2 (2,1 & 2,k)
    - ❑ Each subject is assumed to be assessed by the same group of raters, and these raters were randomly sampled
      - ❑ Still random effect of raters

ICC Models

- Model 3 (3,1 & 3,k)
  - Each subject is assessed by the same group of raters, but these raters are the only ones of interest
    - No desire to generalize the ICCs calculated beyond the confines of the study or laboratory
    - Does not include systematic error in the model

Example

- ❑ Using Model 3,1
  - ❑ Test-retest reliability
  - ❑ No desire to generalize to other devices or testers
  - ❑ Systematic error is not accounted for, but we conduct an ANOVA to test for systematic error
    - ❑ This receives the same criticism as the Pearson R for not accounting for systematic error

Interpreting the ICC

- ❑ If ICC = 0.95
  - ❑ 95% of the observed score variance is due to true score variance
  - ❑ 5% of the observed score variance is due to error

- ❑ 2 factors for examining the magnitude of the ICC
  - ❑ Which version of the ICC was used?
  - ❑ Magnitude of the ICC depends on the between-subjects variability in the data
    - ❑ Because of the relationship between the ICC magnitude and between-subjects variability, standard error of measurement values (SEM) should be included with the ICC

Implications of a Low ICC

❑ Low reliability

❑ Real differences

  ❑ Argument to include SEM values

❑ A low ICC means that more subjects will be necessary to overcome the increased percentage of the observed score variance due to error.