

# 1

# Rehabilitation Engineering

## §1.1 *Rehabilitation outcome measures*

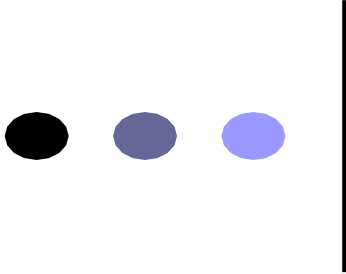
Lorenzo Chiari, PhD

Dipartimento di Ingegneria dell'Energia Elettrica e  
dell'Informazione

***Alma Mater Studiorum – Università di Bologna***

[lorenzo.chiari@unibo.it](mailto:lorenzo.chiari@unibo.it)





*'... when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.'*

(Lord Kelvin, PLA, Vol. 1, Electrical Units of Measurement, 1883-05-03)

# Noise sources

Random  
Systematic

Environment

Measurement

Experimental  
protocol

Physiology



# Noise sources

Random

Systematic

External/Exogenous

Environment

Between  
(Inter-rater)

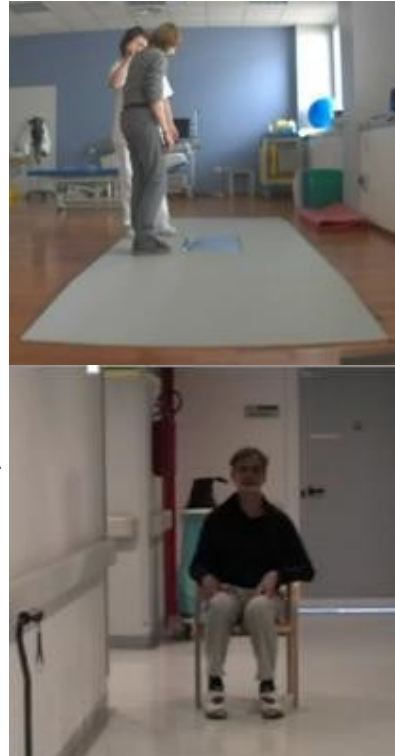
Within  
(Intra-rater)

Experimental  
protocol

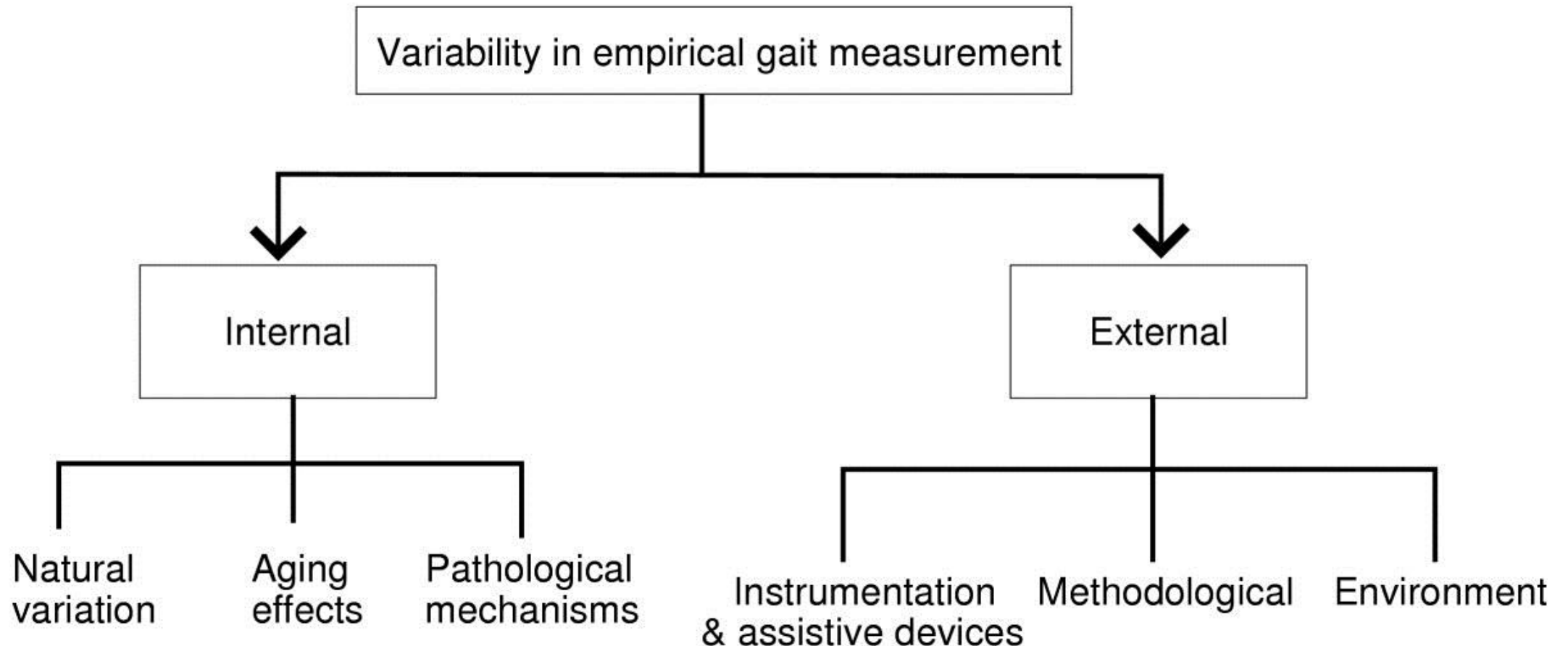
Measurement

Physiology

Internal/Endogenous



# Example: sources of variability in gait measurements





# Quality assessment

- It is impossible to control them all (or at least all those that do not bring the information we are looking for), and then **cancel** them *a priori*. At most, we can try to **minimize** them. A posteriori, then we will have to quantify their residual effects ...
- **How do we check if and to what extent the various noise sources (and related spurious variability) afflict our instrumental measures in the rehabilitation context?**

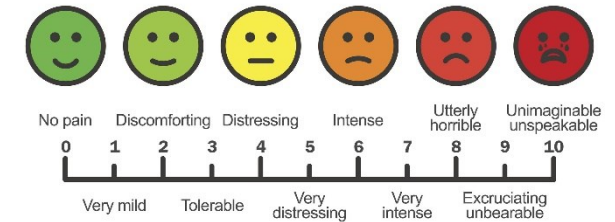
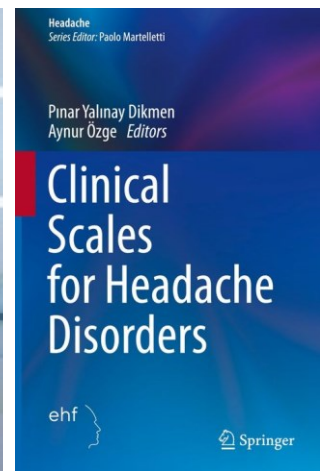
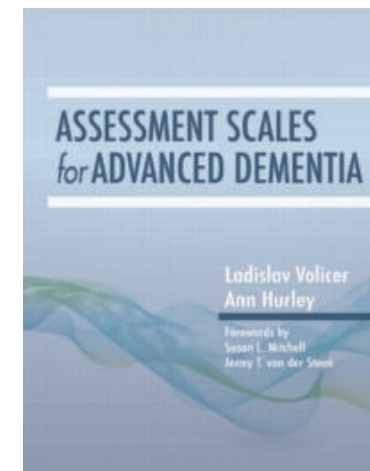


# Measuring, why?

- **To find differences**  
→ Cross-sectional analysis  
[*Diagnosis; differential diagnosis*]
- **To quantify a change**  
→ Longitudinal Analysis  
[*Monitoring; Follow-ups*]
- **To predict the evolution of a system**  
→ Predictive models  
[*Prognosis*]

# Measuring, how?

- Scales are used to manifest latent constructs
- They measure behaviors, attitudes, and hypothetical scenarios we expect to exist due to our theoretical understanding of the world, but cannot assess directly
- Scales are typically used to capture a multifaceted behavior, a feeling, or an action that cannot be captured in a single variable or item
- The use of multiple items to measure an underlying latent construct can additionally account for and isolate item-specific measurement error, which leads to more accurate research findings
- Thousands of scales have been developed that can measure a range of social, psychological, and health behaviors and experiences

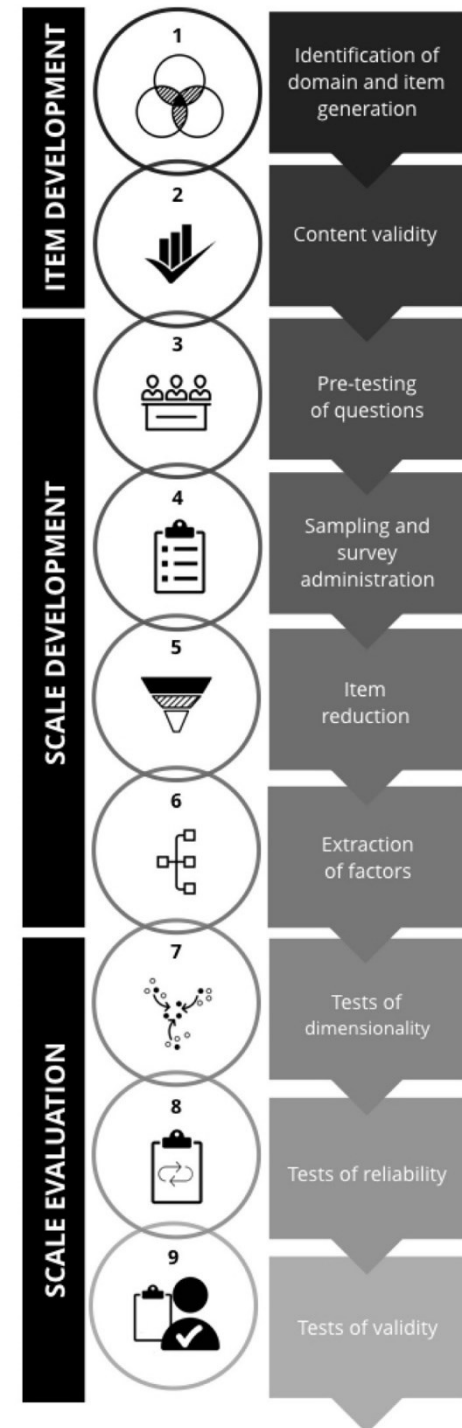




# Measuring, how?

An overview of the three phases and nine steps of scale development and validation

*Boateng et al., Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer, Front Public Health. 2018; 6: 149.*



# The needs

We need measurement tools able to predict:

- the manifestation of the results in the long term,
- their "spontaneous evolution",
- their growth in complexity,
- the global loss of compensatory capacity of the system in relation to the presence of chronic disease.

*(Baratto et al., Problematiche cliniche in riabilitazione, in «Bioingegneria della postura e del movimento», Patron, 2003)*



# Expected measurement properties

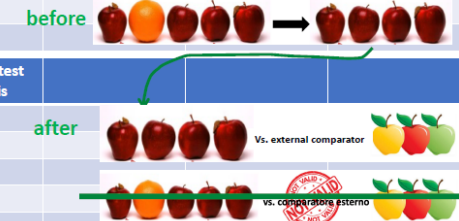
For a measure to be useful, it must provide **accurate** and **meaningful** results to clinicians and researchers:

1. No scaling issues (adequate levels, no floor and ceiling effects)
2. Reliability (relative and absolute):
  - Internal consistency
  - Test-retest
  - Intra-rater
  - Inter-rater
3. Validity
4. Responsiveness
5. Interpretability

(Finch et al., *Physical rehabilitation outcome measures*, 2002)

Measurement models

"Internal" properties	Classical item analysis	Confirmatory Factor analysis	Mokken analysis	IRT 2-3 P models analysis	Rasch analysis
Precision	x			x	x
Targeting	x			x	x
Unidimensionality		x	x	x	x
Linearity		x		x	x
Monotonicity			x	x	x
Subgroup invariance					x
Invariance					x
"External" properties	Classical test analysis				
Concurrent validity	x				
Discriminant validity	x				
Criterion validity	x				
Hypothesis testing	x				
Predictive validity	x				





# Scaling

1

Levels of measurement:

- **Nominal:** represented by categories, no hierarchy among response options (e.g., gender, geographic region);
- **Ordinal:** order is relevant, with hierarchy among response options, but spacing among responses is not viewed as being equal (e.g., level of education);
- **Interval:** response options are equally spaced, however the scale does not have a meaningful zero value (e.g., temperature in °C, HRQOL, most self-report measures)
- **Ratio:** equal spacing between response options and possess a meaningful zero (e.g., temperature in °K)



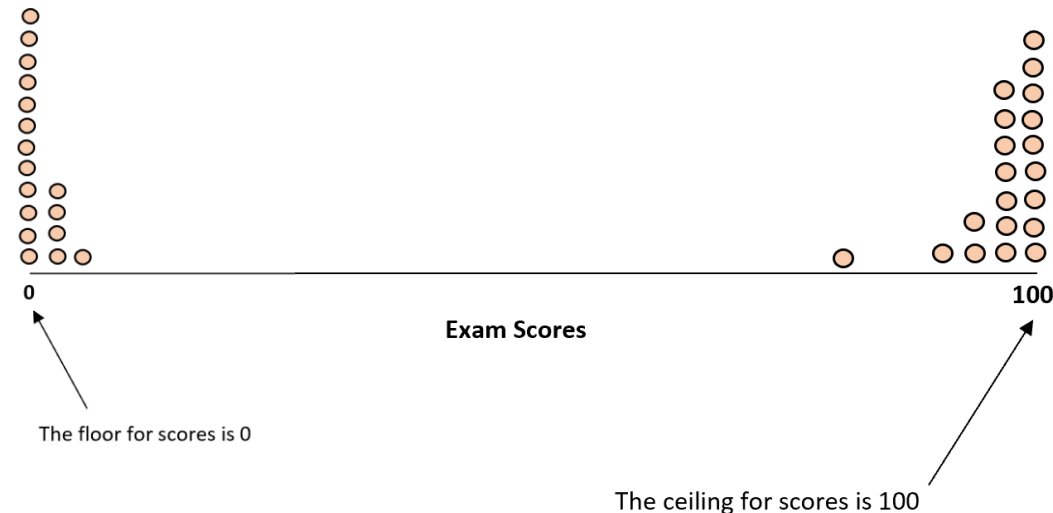
# Scaling

Being able to distinguish among the different levels of measurement is important for two reasons:

- The level of measurement of a scale dictates the mathematical operations that can be performed (e.g., '+' of items only if the scale is interval or ratio; 'x', ':', '%' only if the scale is ratio)
- Reliability and validity indexes are different for different levels (➔ interval or ratio)

# Scaling

A useful measure must provide room on the scale for clients to demonstrate improvement and deterioration → a clinically useful measure must not demonstrate **ceiling** or **floor** effects



A floor/ceiling effect can cause a variety of problems including:

- It makes it difficult to get an accurate measure of central tendency.
- It makes it difficult to get an accurate measure of dispersion.
- It makes it difficult to rank individuals according to score.
- It makes it difficult to compare the means between two groups.



# Reliability

2

A reliable measure must fulfill two requirements:

1. It must provide consistent values with small errors of measurement;
2. It can differentiate among the clients on whom the measurements are being applied

CONSISTENCY

DISCRIMINANT ABILITY

*Example: two handheld dynamometers, one of which is broken*



# Reliability

From these requirements two different methods for describing reliability follow:

**ABSOLUTE RELIABILITY**



**CONSISTENCY**

**RELATIVE RELIABILITY**



**DISCRIMINANT ABILITY**





# Absolute reliability (or agreement)

This term is used to represent the reliability of a measure expressing its measurement error (in the same units as the original measurement).

Absolute reliability is quantified by the **standard error of measurement (SEM)**, not to be confused with the standard error of the mean.

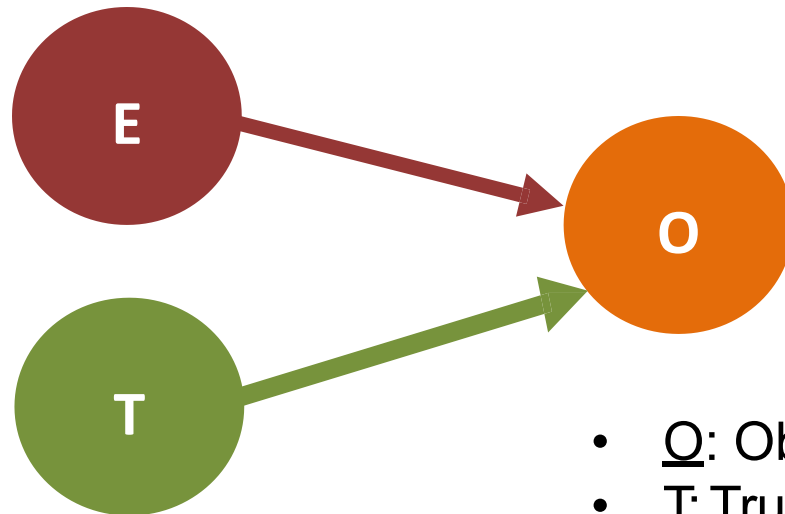
Although there are several algebraically equivalent methods for calculating the SEM, the most conceptually appealing description is that of the square root of the error or within-client variance.

$$SEM = \sqrt{\sigma_{within}^2}$$

# Relative reliability

This term is used to describe a measure's ability to distinguish among clients.

This coefficient is typically based on *Classical Test Theory* (CTT)



- O: Observed score
- T: True score
- E: Error score

$$O = T + E$$

$$O - E = T$$



# Classical Test Theory

- Variance is the statistical term that is used to describe the variability of the data

$$Reliability = \frac{var(T)}{var(O)} \approx \frac{var(O) - var(E)}{var(O)}$$

$$O = T + E$$

$$T = O - E$$

Ideally:  $E \rightarrow 0 \Rightarrow O \approx T \Rightarrow Reliability \approx 1$

**It is an index of how much the variance of the measure is free from the variance of error**



# Classical Test Theory

*Relative Reliability Coefficient =*

$$\frac{\text{var}(T)}{\text{var}(O)} = \frac{\text{true variance}}{\text{true variance} + \text{error variance}}$$

Or, in the clinical setting:

*Relative Reliability Coefficient =*

$$\frac{\text{between client variance}}{\text{between client variance} + \text{within client variance}}$$



# Classical Test Theory

- The *true variance* represents the extent to which clients' average scores differ (between-client variance).
- The *error variance* represents the extent to which replicate measures within a client differ (within-client variance).
- The Relative Reliability Coefficient is unitless. When expressed in this format it is an ***intraclass correlation coefficient***.
- Typically, intraclass correlation coefficients vary from 0 to 1, with higher values representing higher reliability.
- Some have used Pearson's correlation coefficient as a reliability index, but this *does not* represent the true variance divided by the observed variance; rather it describes the association between duplicate measures → not to be used for this analysis.

# Exercise

Suppose that an investigator is interested in assessing the test-retest reliability of the Three-Step Performance Measure. The unit of measurement is distance in meters. Ten patients recovering from ankle sprains were tested on two occasions separated by 48 hours. The data are shown in the Table:

**TABLE 4-4 Three-Step Patient Data Recorded in Meters**

Patient	Time 1	Time 2	Difference
1	16.20	17.90	-1.70
2	5.90	4.40	1.50
3	10.80	8.70	2.10
4	17.20	17.60	-0.40
5	8.00	6.30	1.70
6	12.70	12.80	-0.10
7	4.50	7.20	-2.70
8	19.60	16.60	3.00
9	10.30	12.70	-2.40
10	14.70	15.40	-0.70
Mean	11.99	11.96	0.03
SD	4.99	4.99	1.98

# Exercise

The results from the analysis of variance of the data are reported in the following table:

**TABLE 4-5 Analysis of Variance Table and Variance Component Calculations**

Source	DF	SS	MS	Variance Components
Between clients	9	430.62	47.85	$\sigma_B^2 = \frac{MSB - MSW}{k} = \frac{47.85 - 1.76}{2} = 23.05$
Within clients	10	17.57	1.76	$\sigma_W^2 = MSW = 1.76$
Total	19			$\sigma_B^2 + \sigma_W^2 = 24.81$

k is equal to the number of measurements on a patient.

Absolute reliability:  $SEM = \sqrt{\sigma_W^2} = 1.33 \text{ m}$

Relative reliability:  $R = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = 0.93$



# Types of reliability

- **Internal consistency:** *based on parallel assessments of clients at an instant in time*, provides an index of a test's ability to differentiate among clients at an instant in time. Applicable when multi-item measures are summarized into a single score. Used metrics: split-half coefficient; coefficient  $\alpha$  (ICC(3,k)). Range from 0 to 1, higher values represent higher levels of internal consistency.
- **Test-retest reliability:** *based on parallel assessments of clients on different occasions*. Provide information about the stability of clients' responses over time. Typical study design involves two or more assessments over an interval when clients are believed to be stable with respect to the attribute of interest. Used metrics: ICC.
- **Interrater reliability:** *based on parallel assessments by different raters*. Of interest when raters are part of the measurement process. Used metrics: ICC.





# Reliability Analysis: tools

We want to quantify and compare the possible contributions to the variance of the measurements made.

The variance of the measures is divided into variance due to the belonging to a particular category of the only independent variable (variance *between*), and to other causes, including error (variance *within*).

The variance *between* is that due to the differences in the measurements between the units of analysis (subjects), while the variance *within* is that due to the random measurement error, regardless of the variations in the true measurements.



## One-way ANOVA

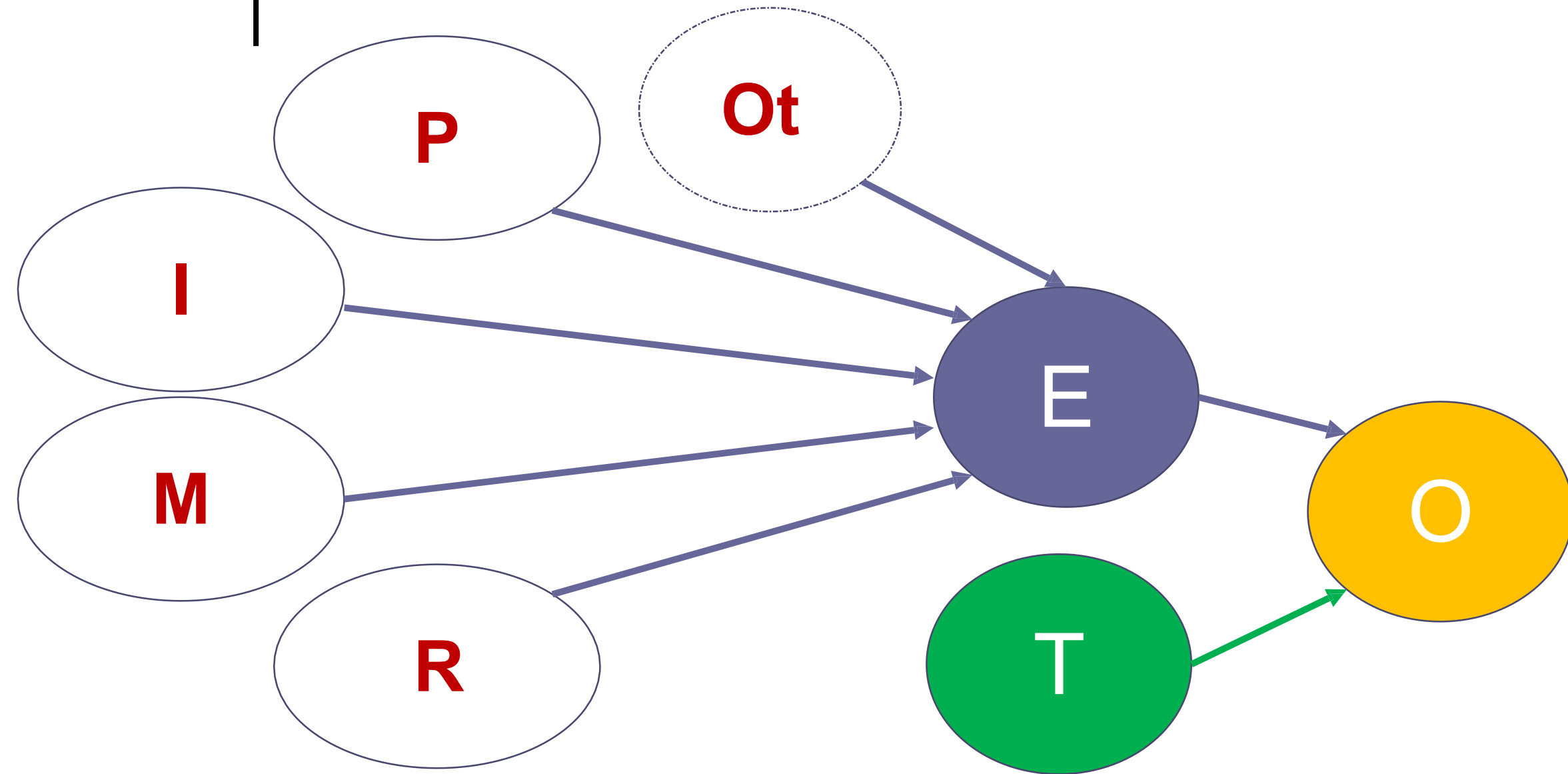


# Reliability Analysis: tools

- By definition, the reliability is included in the interval  $[0 \ 1]$ .
- It is measured through the correlation between different instances of the same variable (it is therefore a univariate statistic) which produces a so-called intra-class correlation coefficient (ICC).
- There are different formulas for the ICC that can give different results if applied to the same data. Each formula is appropriate for specific situations that are defined by the experimental design and the potential use to be made of the results.
- Virtually, **all the ICC values can be calculated starting from an analysis of variance (ANOVA) carried out on the data matrix.**



...but reality is far more complex...





# Main sources of variance inflation

**P**articipant (measurand): mood, motivations, fatigue, memory, health condition, fluctuations, previous practice, specific knowledge and familiarity with the test, ...

**I**nstructions and tests (protocol / procedure): clarity and completeness of the instructions, adherence to instructions, consistency in providing instructions by the evaluator, ...

**M**easurement (measuring instruments): lack of calibration, electronic, mechanical, thermal noise, errors in the measurement model or inadequacy of the chosen instrument in relation to the test, the evaluator and the participant, ...

**R**ater: competence, experience and concentration of the evaluator in relation to the complexity of the assessment, familiarity with the test to be assessed

**O**ther: other possible sources of noise not included in the previous ones



# A more general model

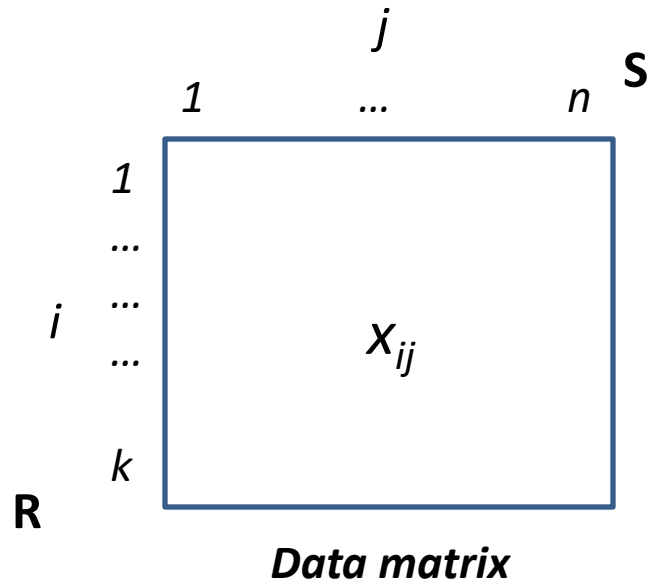
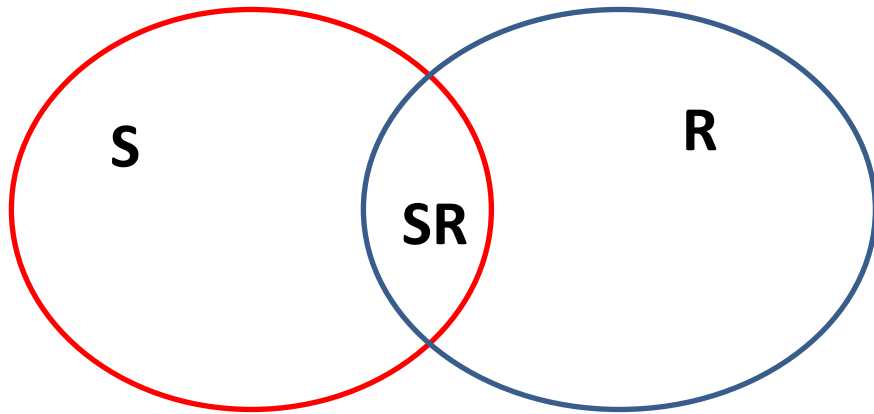
- The **theory of generalizability** (TG, or Generalizability Theory, Cronbach et al., 1972) offers a broad frame of reference for examining in more depth the reliability of the measures.
- TG extends the classical concept of reliability of classical test theory (TCT) by **trying to identify and quantify the contribution of more than just one source of measurement error**.
- The measurement error, as known, can indeed derive from a multiplicity of sources, so the estimate of the error variance and therefore the reliability of the test vary depending on the way in which the data is collected.



# A more general model

- The TG tries to go beyond the simple investigation of the accuracy of the observed score compared to the true score, as it asks at what level of accuracy the observed score makes it possible to generalize people's behavior to a universe of situations.
- From a statistical point of view, the TG is based on the Analysis of Variance (ANOVA), as **it considers the possible sources of measurement error as factors (or independent variables) in a factor variance analysis model.**
- Factorial ANOVA allows for the introduction of multiple causes in the model for explaining the variance of the observed score, which not only contribute to this variance independently of each other (the so-called main effects) but also interacting with each other, hence the interaction effects.
- **The TG assumes that the causes of measurement errors are multiple, so it uses a consistent statistical frame of reference.**

# A more general model



Example:  
*S* – subjects  
*R* – repetitions

ANOVA1:  $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

ANOVA2:  $Y_{ijw} = \mu + \alpha_i + \beta_j + \varphi_{ij} + \varepsilon_{ijw}$

ANOVA\_RM:  $Y_{ij} = \mu + \pi_i + \tau_j + \pi\tau_{ij} + \varepsilon_{ij}$

Labels and annotations:

- $\mu$ : Grand average (population mean)
- $\alpha_i$ : Effect due to Factor 1 (e.g., a treatment)
- $\varepsilon_{ij}$ : Residuals (random error)
- $\alpha_i$ ,  $\beta_j$ ,  $\pi_i$ ,  $\tau_j$ : Effects due to Factors 1 and 2
- $\varphi_{ij}$ ,  $\pi\tau_{ij}$ : Interaction effect between Factors

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

# ANOVA models used for ICC

Table 3  
Mean Square Expectations for Analysis of Variance Models Given in Table 1

Model and source of variation	<i>df</i>	<i>MS</i>	<i>EMS</i>
Case 1: One-way random effects model			
Between rows	$n - 1$	$MS_R$	$k\sigma_r^2 + \sigma_w^2$
Within rows	$n(k - 1)$	$MS_W$	$\sigma_w^2$
Case 2: Two-way random model with interaction			
Between rows	$n - 1$	$MS_R$	$k\sigma_r^2 + \sigma_c^2 + \sigma_e^2$
Within rows	$n(k - 1)$	$MS_W$	$\sigma_r^2 + \sigma_c^2 + \sigma_e^2$
Between columns	$k - 1$	$MS_C$	$n\sigma_c^2 + \sigma_r^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	$MS_E$	$\sigma_r^2 + \sigma_e^2$
Case 2A: Two-way random model, interaction absent			
Between rows	$n - 1$	$MS_R$	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	$MS_W$	$\sigma_r^2 + \sigma_e^2$
Between columns	$k - 1$	$MS_C$	$n\sigma_c^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	$MS_E$	$\sigma_e^2$
Case 3: Two-way mixed model with interaction			
Between rows	$n - 1$	$MS_R$	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	$MS_W$	$\theta_c^2 + \frac{k}{k - 1}\sigma_c^2 + \sigma_e^2$
Between columns	$k - 1$	$MS_C$	$n\theta_c^2 + \frac{k}{k - 1}\sigma_c^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	$MS_E$	$\frac{k}{k - 1}\sigma_c^2 + \sigma_e^2$
Case 3A: Two-way mixed model, interaction absent			
Between rows	$n - 1$	$MS_R$	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	$MS_W$	$\theta_c^2 + \sigma_e^2$
Between columns	$k - 1$	$MS_C$	$n\theta_c^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	$MS_E$	$\sigma_e^2$

Note.  $E(MS)$  = expected mean squares;  $MS_R$  = mean square for rows;  $MS_W$  = mean square for residual sources of variance;  $MS_C$  = mean square for columns;  $MS_E$  = mean square error.

McGraw and Wong, 1996



# ICCs for One-Way and Two-Way ANOVAs

McGraw and Wong, 1996

Table 4  
Single Score Intraclass Correlation Coefficients (ICCs) for One-Way and Two-Way Models

Definitions of ICCs $\rho$	Formulas for calculating $\hat{\rho}$	Designation	Interpretation of ICC
Row effects random			
One-way model Case 1 model $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_w^2}$	$\frac{MS_R - MS_W}{MS_R + (k-1)MS_W}$	ICC(1)	The degree of absolute agreement among measurements made on randomly selected objects. It estimates the correlation of any two measurements.
Column and row effects random (two-way random effects model)			
Two-way models <sup>a</sup> Case 2 model $\frac{\sigma_i^2}{\sigma_i^2 + (\sigma_c^2 + \sigma_e^2)}$ or Case 2A model $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_c^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$	ICC(C,1)	The degree of consistency among measurements. Also known as norm-referenced reliability and as Winer's adjustment for anchor points (Winer, 1971). In generalizability theory, this ICC estimates the squared correlation of individual measurements and universe scores.
Case 2 model $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_c^2 + (\sigma_n^2 + \sigma_e^2)}$ or Case 2A model $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_c^2 + \sigma_n^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$	ICC(A,1)	The degree of absolute agreement among measurements. Also known as criterion-referenced reliability. Estimates the Type 1 ICC for one-way, unmatched data (Rajaratnam, 1960).

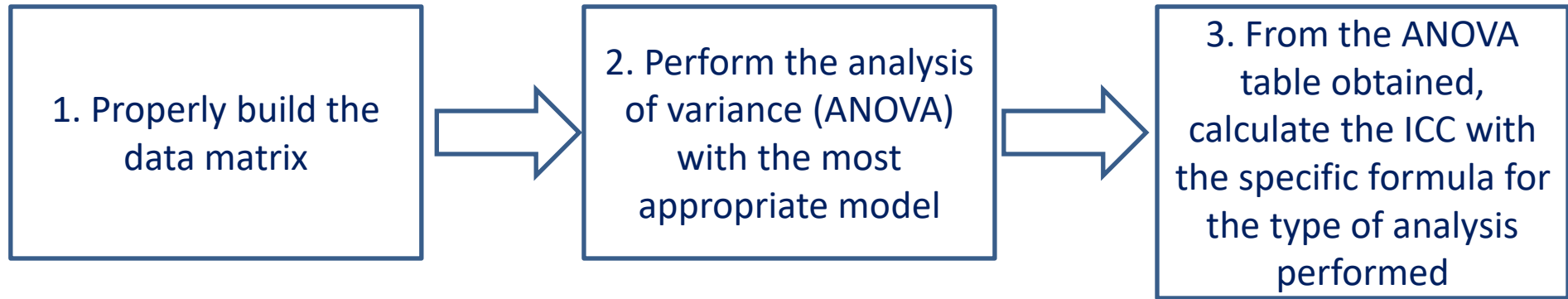
Column effects fixed, row effects random (two-way mixed effect model)			
Case 3 model $\frac{\sigma_i^2 - \sigma_n^2/(k-1)}{\sigma_i^2 + (\sigma_c^2 + \sigma_e^2)}$ or Case 3A model $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$	ICC(C,1)	The degree of consistency among measurements made under the fixed levels of the column factor. This ICC estimates the correlation of any two measurements, but when interaction is present, it underestimates reliability.
Case 3 model $\frac{\sigma_i^2 - \sigma_n^2/(k-1)}{\sigma_i^2 + \theta_c^2 + (\sigma_n^2 + \sigma_e^2)}$ or Case 3A model $\frac{\sigma_i^2}{\sigma_i^2 + \theta_c^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$	ICC(A,1)	The absolute agreement of measurements made under the fixed levels of the column factor.

Note.  $MS_R$  = mean square for rows;  $MS_W$  = mean square for residual sources of variance;  $MS_E$  = mean square error;  $MS_C$  = mean square for columns.

<sup>a</sup> In the event of data with a two-way classification for which the column variance is zero (i.e.,  $\sigma_c^2 = 0$  or  $\theta_c^2 = 0$ , depending on the model), a one-way model should be used. Thus even though test scores on  $k$  parallel tests can be classified by test and test taker, the column variance by definition is zero, which means that a one-way model applies.



# General workflow for computing the ICC

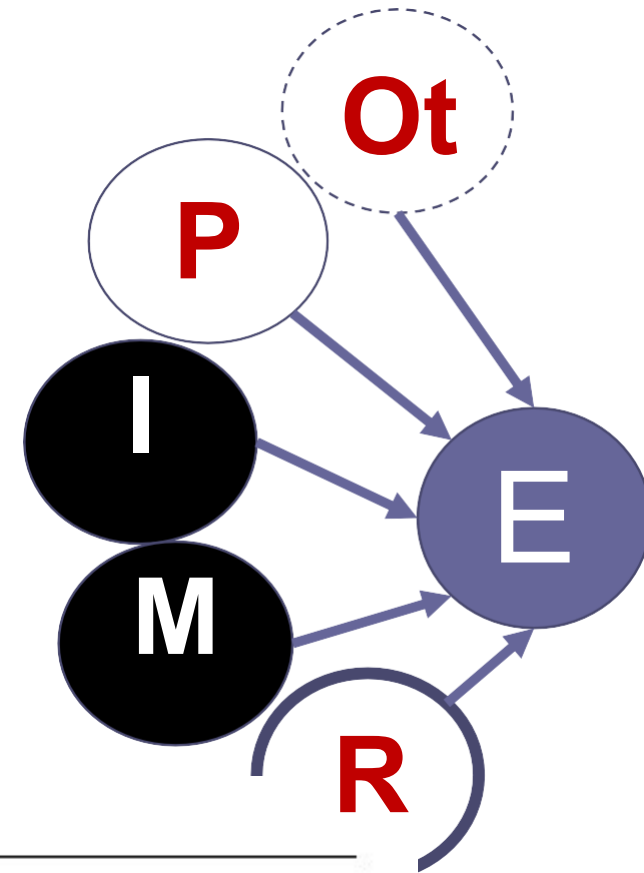


# CASE 1: Intra-rater reliability

- Number of raters: **1**
- Number of measures taken by the same rater: **k**
- Number of analyzed subjects: **n**

## Procedure:

1. One-Way ANOVA (random effects model)
2. Compute ICC(1,1) with the following formula



**Table 3** ANOVA table of results for intra-rater reliability – one-way random effects model

Source of variation	Degrees of freedom (df)	Mean square (MS)	Expected mean square E(MS)
Between-subjects	$n-1$	BMS	$\sigma_e^2 + k\sigma_s^2$
Within-subjects	$n(k-1)$	WMS	$\sigma_e^2$

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1) WMS}$$

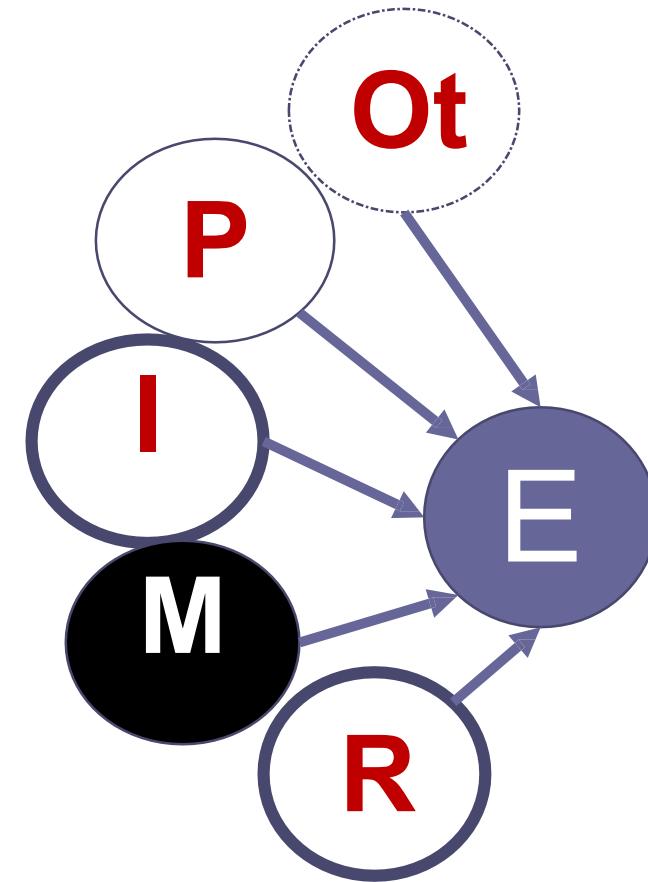
**k** is the number of times; **n** is the number of subjects;  $\sigma^2$  = variance.

# CASE 2: Inter-rater reliability

- Number of raters:  $k$  (*the exclusive raters of interest*)
- Number of measures taken by the same rater:  $1$
- Number of analyzed subjects :  $n$

## Procedure:

1. Two-Way ANOVA (mixed model, raters fixed)
2. Compute ICC(3,1) with the following formula



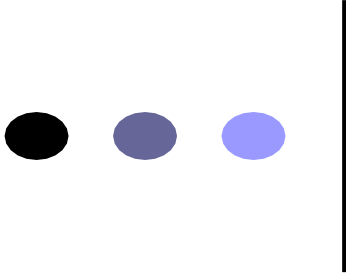
**Table 2** ANOVA table of results for inter-rater reliability – two-way mixed model, raters fixed

Source of variation	Degrees of freedom (df)	Mean square (MS)	Expected mean square E(MS)
Between-subjects	$n-1$	BMS	$\sigma_e^2 + k\sigma_s^2$
Between-raters	$k-1$	RMS	$\sigma_e^2 + n/k-1 \sum \rho_j^2$
Error	$(n-1)(k-1)$	EMS	$\sigma_e^2$

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1) EMS}$$

$k$  is the number of raters;  $n$  is the number of subjects;  $\sigma^2$  variance. For expected mean square equations see Fleiss.<sup>11</sup>

In this case, the analysis measures the agreement between the raters, who are all the observers of interest. The result is not generalizable to other observers.



## CASE 3: Inter-rater reliability (or objectivity)

- Number of raters:  $k$  (*taken from a wider set of raters to which we want to generalize the result obtained*)
- Number of measures taken by the same rater:  $1$
- Number of analyzed subjects :  $n$

### ***Procedure:***

1. Two-Way ANOVA (mixed model)
2. Compute ICC(2,1) with the following formula

$$\text{ICC (2,1)} = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k-1)\text{EMS} + k(\text{RMS}-\text{EMS})/n}$$

In this case, the analysis measures the agreement between the raters, but also their interchangeability.



# How to interpret the ICC value?

- There are no universal and shared cut-off values

Proposal by Fleiss (1986):

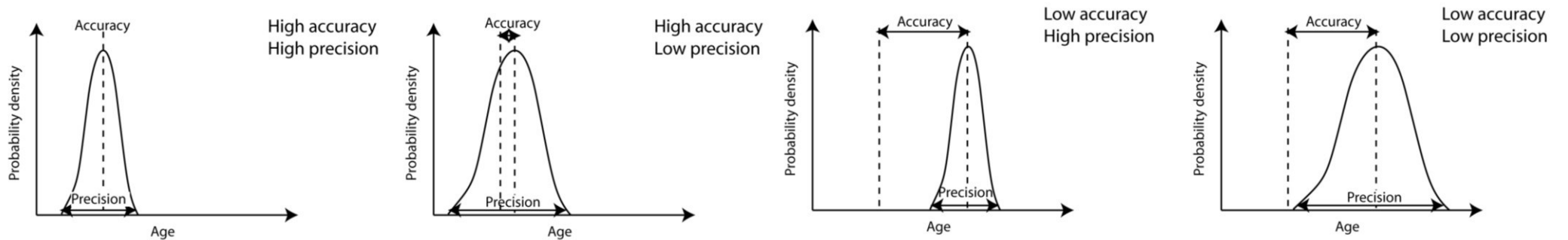
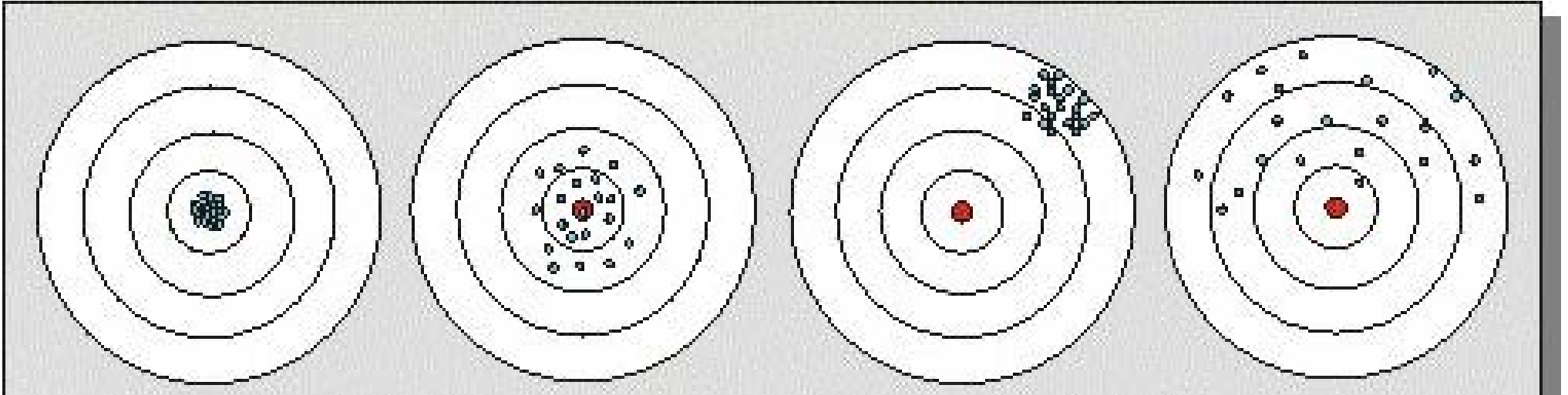
- |                      |   |                                 |
|----------------------|---|---------------------------------|
| • $ICC \geq 0,75$    | - | <i>excellent</i> reliability    |
| • $0,4 < ICC < 0,75$ | - | <i>fair-to-good</i> reliability |
| • $ICC \leq 0,4$     | - | <i>poor</i> reliability         |



# Reliability in different types of studies

- **Evaluation study:** the evaluation of the same case (same quantity, measured in the same subject in the same conditions) must give the same outcome → absolute reliability is important
- **Discriminative study:** patients under the same conditions must receive the same assessment, regardless of the evaluator (rater), the moment and the context → relative reliability is important
- **Reliability is a necessary condition but not sufficient for validity:** you cannot trust a measure/test that provides different outputs on repeated tests in sequence or using different measurement / evaluator tools, or which has no discriminatory capacity, etc., However, a repeatable measure/test may not be valid.

# Validity and reliability



**Valid measures**

**Reliable**

**Unreliable**

**Non valid measures**

**Reliable**

**Unreliable**





# Validity

3

## Validity:

- **A measure / instrument is valid to the extent that it assesses what it is intended to measure**
  - Accordingly, validity is not an “all or none” property but rather a matter of degree.
  - Knowledge of a measure’s validity is constantly evolving as new information becomes available.
  - Traditionally, validity has been divided into the following topics:
    - ***Face validity***
    - ***Content validity***
    - ***Criterion validity***
    - ***Construct validity***
- More recently, additional terms have been applied



# Content-related Validity

**Face validity:** considers whether a measure appears to be measuring what it is intended to measure.

*Example: if the goal of a measure is to assess lower extremity functional status, one would expect to see items/measurements concerning walking, standing, running, and negotiating stairs. However, if a measure designed to assess lower extremity functional status focused primarily on upper extremity tasks or inquired about emotional well-being, one would question its face validity.*

**Content validity:** refers to how well a measure/test measures the construct that it sets out to measure. It exists to the extent that a measure is composed of a comprehensive sample of items/measurements that precisely and completely assess the domain of interest.

*Examples: 1) If the goal is to measure lower extremity functional status, one would expect a set of activities that cover all aspects of lower extremity function. Simply measuring a client's walking distance in 2 minutes or asking them about their ability to climb a flight of stairs is not a comprehensive set of lower extremity items/measurements: they sample only a select aspect of the spectrum of activities associated with lower extremity function.*

*2) Suppose a professor wants to test the overall knowledge of his students in the subject of ageing and rehabilitation engineering. His test would have content validity if:*

- The test covers every topic of the course that he taught in the class.*
- The test does not cover unrelated topics such as history, economics, biology, etc.*

*A test lacks content validity if it doesn't cover all aspects of a construct it sets out to measure or if it covers topics that are unrelated to the construct in any way.*



# Criterion-related Validity

**Criterion validity:** examines the extent to which a measure provides results that are consistent with a gold standard.

*Example: if the goal is to determine the validity of a manual muscle testing, the results from a manual muscle test could be compared with results from an assessment using a dynamometer that does not involve the assessor to physically interact with the client. In this case the dynamometer's results represent the gold standard.*

Typically, criterion validity is divided into:

- **Concurrent validity:** compares the measure's results to the gold standard's results that is obtained at (approximately) the same point in time.

*Example: see previous example*

- **Predictive validity:** examines a measure's ability to predict some subsequent criterion event.

*Example: use of the Berg Balance Test or the gait speed to predict falls over the following 6 weeks. In this case, the criterion standard would be whether the patient fell over the next 6 weeks.*



# Construct-related Validity

**Construct validity:** for some attributes (e.g., pain, health-related quality of life), no criterion standard exists. It is usually insufficient to depend on face and content validity alone. In the absence of a gold standard, a construct validation process is applied. It involves forming theories about the attribute of interest and then assessing the extent to which the measure under investigation provides results that are consistent with the theories.

*Example: several theories applied to the construct validation of functional status measures in clients with low back pain (→ “known group difference method”). E.g., clients presenting with acute low back pain are more disabled than clients with longer-standing episodes of back pain; clients who have low back pain and pain radiating into the lower extremity are more disabled than clients who present with back pain only. In such cases, the measure will have construct validity if it will be able to capture properly these expected differences.*

A more recent interpretation of construct validation is that it is not restricted to testing theories; rather, it includes all aspects of validity (i.e., face, content, and criterion).



# Comparative Validity

**Cross-sectional and longitudinal validity:** these terms have been applied within the realm of construct validation to distinguish between measures taken at a single point in time (*cross-sectional*) and measures that relate to change scores (*longitudinal*).



**Convergent validity:** examines the extent to which a measure's result agrees with the result of another measure that is believed to be assessing the same attribute. Correlation coefficients are often used to quantify the convergent validity of a measure. If the comparison measure is the gold standard, this would represent criterion validity.

*Cross-sectional example: the results from the measure of interest obtained at a single point in time should demonstrate a moderately high correlation with the results from a second measure that has been validated previously for the same purpose.*

*Longitudinal example: the assessment of change by the measure of interest should demonstrate a moderately high correlation with the results of change from a second measure that has been validated previously for the same purpose.*



# Comparative Validity

**Cross-sectional and longitudinal validity:** these terms have been applied within the realm of construct validation to distinguish between measures taken at a single point in time (*cross-sectional*) and measures that relate to change scores (*longitudinal*).

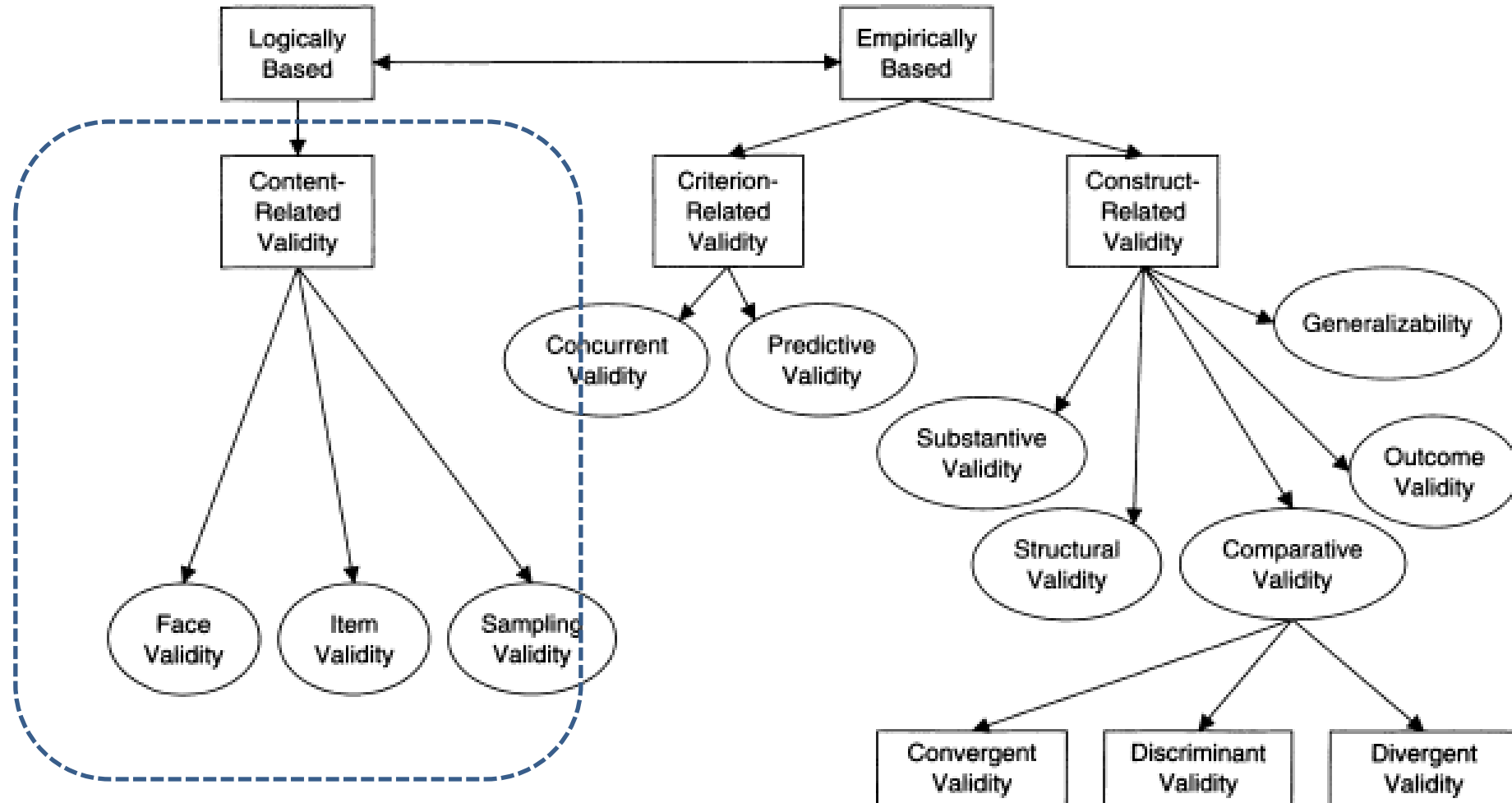


**Discriminant validity:** examines the extent to which a measure *correlates* with measures of attributes that are different from the attribute the measure is intended to assess. Correlation coefficients are often used to quantify the discriminant validity of a measure. One would expect a measure designed for a specific purpose to perform better when assessing the attribute of interest than it would if used to assess other attributes. If a measure correlated highly with a spectrum of attributes, a possible interpretation is that the measure is assessing a general concept (e.g., overall well-being) rather than the specific attribute of interest.

*Cross-sectional example: clients' scores on a lower extremity functional status measure will correlate more highly with the physical functional subscale scores from a generic health status measure than with the emotional subscale scores from the generic measure.*

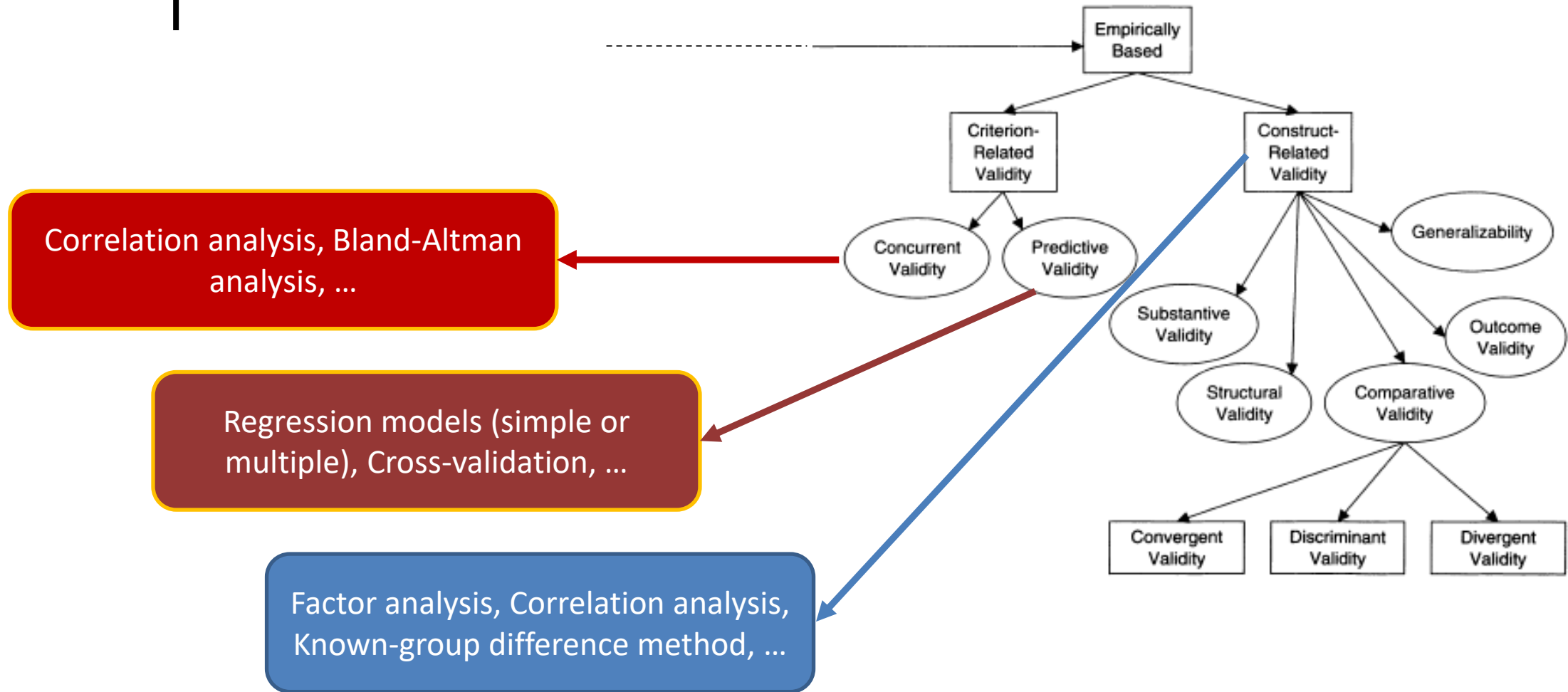
*Longitudinal example: clients' change scores on a lower extremity functional status measure will correlate more highly with the physical functional subscale change scores from a generic measure than with the emotional subscale change scores from the generic measure.*

# Different types of validity



Analysis not supported by any statistical evidence

# Validity analysis: statistical tools







# Bland-Altman method

**AIM: assess the agreement or concurrent validity between two measurement methods**

- In 1983 Altman and Bland (B&A) proposed an analysis, based on the quantification of the agreement between two quantitative measurements by studying the *mean difference* and constructing *limits of agreement*.
- The B&A plot analysis is a simple way to evaluate a *bias* between the mean differences, and to estimate an agreement interval, within which 95% of the differences of the second method, compared to the first one, fall. Data can be analyzed both as unit differences plot and as percentage differences plot.
- The B&A plot method only defines the intervals of agreements, it does not say whether those limits are acceptable or not. Acceptable limits must be defined a priori, based on clinical necessity, biological considerations or other goals.



# Bland-Altman method

- They established a method to quantify agreement between two quantitative measurements by constructing limits of agreement. These statistical limits are calculated by using the mean ( $m$ ) and the standard deviation ( $sd$ ) of the differences between two measurements A and B. To check the assumptions of normality of differences and other characteristics, they used a graphical approach.
- The resulting graph is a scatter plot XY, in which the Y axis shows the difference between the two paired measurements (A-B) and the X axis represents the average of these measures  $((A+B)/2)$ . In other words, the difference of the two paired measurements is plotted against the mean of the two measurements. B&A recommended that 95% of the data points should lie within  $\pm 2sd$  of the mean difference.



# Bland-Altman plot

## Algorithm

1. Calculate the means and differences between the two corresponding readings →  $MEAN = (A+B)/2$  and  $DIFF = A-B$
2. Plot MEAN on the X axis and DIFF on the Y axis of a scatterplot
3. Calculate the mean ( $m_{DIFF}$ ) and the standard deviation ( $sd_{DIFF}$ ) of vector DIFF
4. Compute the limits of agreement as\*

$$LOA = m_{DIFF} \pm 2sd_{DIFF}$$

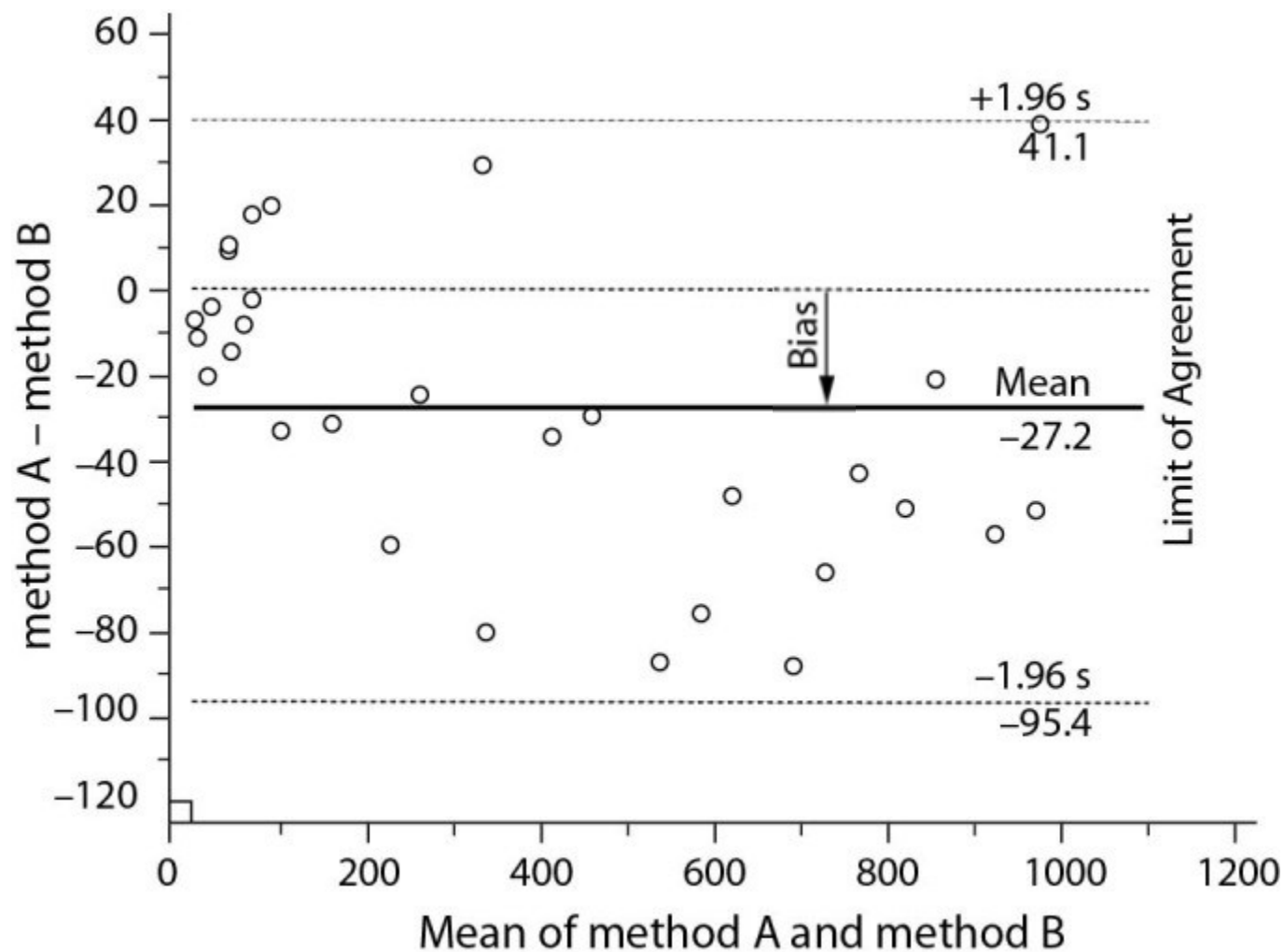
### Agreement indicators:

- $bias = m_{DIFF} \approx 0$
- $sd_{DIFF} \ll$

\*We expect most of the differences to lie between  $m_{DIFF} - 2sd_{DIFF}$  and  $m_{DIFF} + 2sd_{DIFF}$ , or more precisely, 95% of differences will be between  $m_{DIFF} - 1.96sd_{DIFF}$  and  $m_{DIFF} + 1.96sd_{DIFF}$ , if the differences are normally distributed (Gaussian).

*Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10. PMID: 2868172.*

# Example



Hypothetical data of an agreement between two methods (Method A and B).

Method A (units)	Method B (units)	Mean (A+B)/2 (units)	(A - B) (units)	(A - B)/ Mean (%)
1.0	8.0	4.5	-7.0	-155.6%
5.0	16.0	10.5	-11.0	-104.8%
10.0	30.0	20.0	-20.0	-100.0%
20.0	24.0	22.0	-4.0	-18.2%
50.0	39.0	44.5	11.0	24.7%
40.0	54.0	47.0	-14.0	-29.8%
50.0	40.0	45.0	10.0	22.2%
60.0	68.0	64.0	-8.0	-12.5%
70.0	72.0	71.0	-2.0	-2.8%
80.0	62.0	71.0	18.0	25.4%
90.0	122.0	106.0	-32.0	-30.2%
100.0	80.0	90.0	20.0	22.2%
150.0	181.0	165.5	-31.0	-18.7%
200.0	259.0	229.5	-59.0	-25.7%
250.0	275.0	262.5	-25.0	-9.5%
300.0	380.0	340.0	-80.0	-23.5%
350.0	320.0	335.0	30.0	9.0%
400.0	434.0	417.0	-34.0	-8.2%
450.0	479.0	464.5	-29.0	-6.2%
500.0	587.0	543.5	-87.0	-16.0%
550.0	626.0	588.0	-76.0	-12.9%
600.0	648.0	624.0	-48.0	-7.7%
650.0	738.0	694.0	-88.0	-12.7%
700.0	766.0	733.0	-66.0	-9.0%
750.0	793.0	771.5	-43.0	-5.6%
800.0	851.0	825.5	-51.0	-6.2%
850.0	871.0	860.5	-21.0	-2.4%
900.0	957.0	928.5	-57.0	-6.1%
950.0	1001.0	975.5	-51.0	-5.2%
1000.0	960.0	980.0	40.0	4.1%
mean ( $\bar{x}$ )			-27.17	-17.40%
standard deviation (s)			34.81	-12.64%



# Example

From our example, the average of the differences is -27.17 units (bottom line of the [table](#)). This mean difference ( $d$ ) is not zero, and this means that on average the second method (B) measures 27.17 units more than the first one. This bias could be a constant or an average result arising from problems for specific concentrations or values. It is important to evaluate the differences at different magnitudes of the measured variable. If neither of the two methods is a “reference”, the differences could be compared with the mean of the two paired values. The average can be seen in column 3. The B&A graph plot simply represents every difference between two paired methods against the average of the measurement, as shown in [Figure](#). The differences between method A and method B are plotted against the mean of the two measurements. *Plotting difference against mean also allows us to investigate any possible relationship between measurement error and the true value. But since we do not know the true value, the mean of the two measurements is the best estimate we have.* If the first method is a standard or reference method, we can use these values instead of the mean of the two measurements, although this is controversial, because a plot of the difference against a “standard measurement” will always appear to show a relation between difference and magnitude when there is none.



# Example

The bias of -27.2 units is represented by the gap between the X axis, corresponding to zero differences, and the parallel line to the X axis at -27.2 units. This negative bias seems to be due to measurements over 200 units, while for lower concentrations data are closer to each other. A negative trend seems to be evident along the graph. Drawing a regression line of the differences could help in detecting a proportional difference. The visual examination of the plot allows us to evaluate the global agreement between the two measurements.



# Responsiveness

4

Responsiveness is defined as «*the ability of an instrument to measure a meaningful or clinically important change in a clinical state. It implies a change that is noticeably, appreciably different that is of value to the patient or physician. The change may allow the individual to perform some essential task or to perform tasks more efficiently or with less pain or difficulty. These changes also should exceed variation that can be attributed to chance.*» (Liang, 2000)

# Responsiveness

- The classification system describes many different categories of change that can be quantified in studies of responsiveness.
- Each category is defined by the place it occupies on a triaxial matrix in which the three axes define the **Who**, **Which**, and **What** of the change being quantified within a study of responsiveness.

**The Who Axis: Who are the results presented for? Individual-level versus Group-level of analysis and interpretation.**

1. Group-level interpretation
2. Individual-level interpretation

**The Which Axis: Which scores are being contrasted?**

1. Between person differences at one point in time
2. Within person change over time
3. Hybrid/both 1 & 2: between person differences of within-person change.

**The What Axis: What type of change being quantified in study**

- i. Minimum potentially detectable change by the instrument
- ii. Minimum change detectable given the measurement error of the instrument.
- iii. Observed change measured by the instrument in a given population.
- iv. Observed change in a population deemed to have improved
  - by patient
  - by clinician/researcher
  - by payer
  - by society
- v. Observed change in those deemed to have had an important improvement
  - by patient
  - by clinician/researcher
  - by payer
  - by society

**Taxonomy of change in studies of responsiveness**

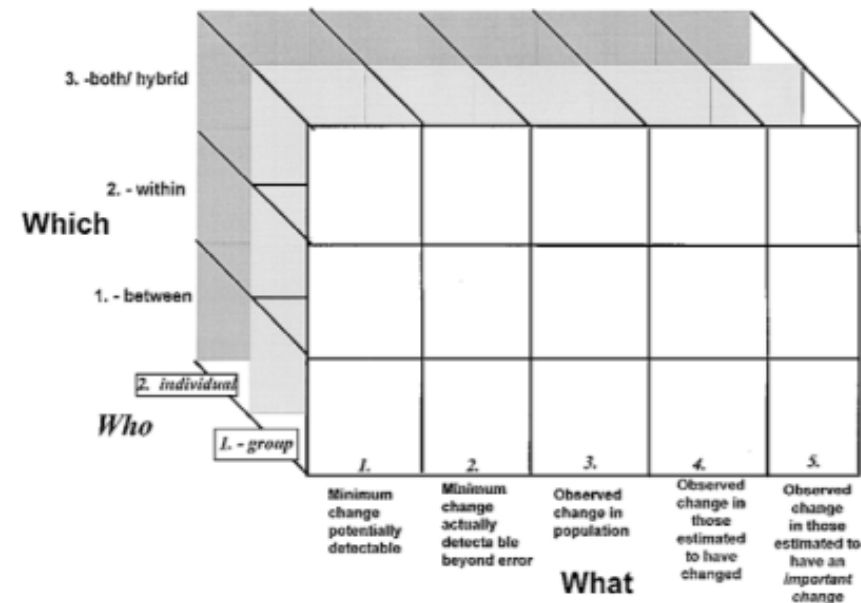


Fig. 5. Depiction of the recommended taxonomy for responsiveness. Each cell in this three-dimensional structure (as defined by the axes of “which, who and what” of the change being quantified), represents a category or type of responsiveness. A given study usually is looking at one kind of change, and therefore one category of responsiveness. A given measure may have information in terms of its responsiveness to one category of change, but this may not mean it is responsive to another.



# Responsiveness

ii. *Minimum change detectable given the measurement error of the instrument (“minimally detectable change”)*

- The second type of change is defined by the error associated with the measurement. Measurement of change reflects true change plus error. When the error is great, wider confidence intervals apply to the observed change score. It can be estimated by:

$$MDC = 1.96 \cdot SEM \cdot \sqrt{2}$$

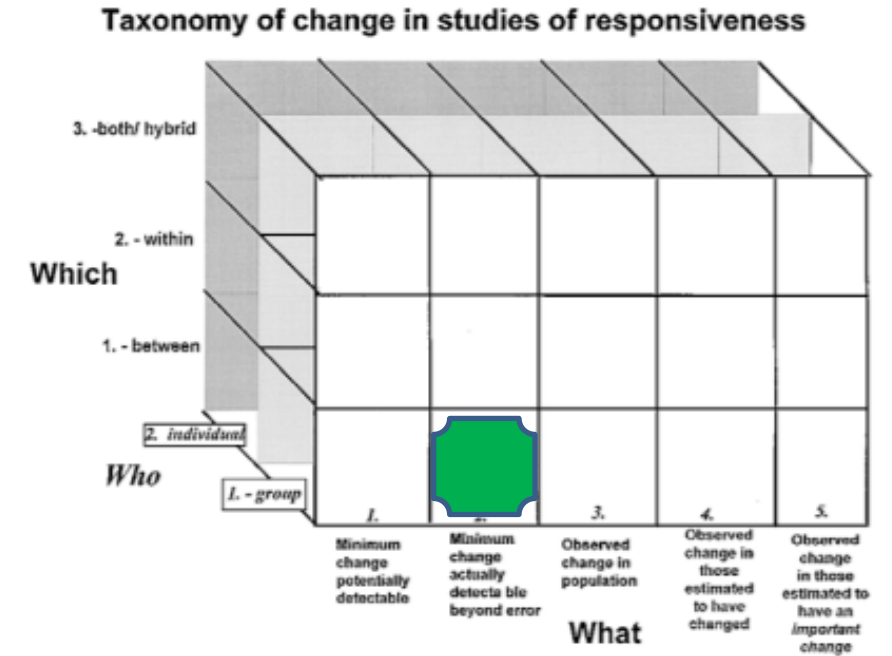


Fig. 5. Depiction of the recommended taxonomy for responsiveness. Each cell in this three-dimensional structure (as defined by the axes of “which, who and what” of the change being quantified), represents a category or type of responsiveness. A given study usually is looking at one kind of change, and therefore one category of responsiveness. A given measure may have information in terms of its responsiveness to one category of change, but this may not mean it is responsive to another.

# Responsiveness

v. *Observed change measured by an instrument in a population deemed to have had an important improvement or deterioration (“Important change”)*

- Important change is estimated change that **is seen to be valued or important** by someone (e.g., patient, clinician, payer) and becomes a criterion for stratifying a sample prior to analysis.
- Subjects who undergo important change (usually separating improvement from deterioration) would be used in the analysis of this type of responsiveness.
- It is often assumed that the *magnitude* and *importance* of change are correlated, but they are not the same. For instance, some research suggests that the threshold for change to be described as “important” varies according to the severity of condition at the time of the baseline assessment.
- The MCID can be calculated using consensus (e.g., Deplhi), anchor, and distribution-based methods.

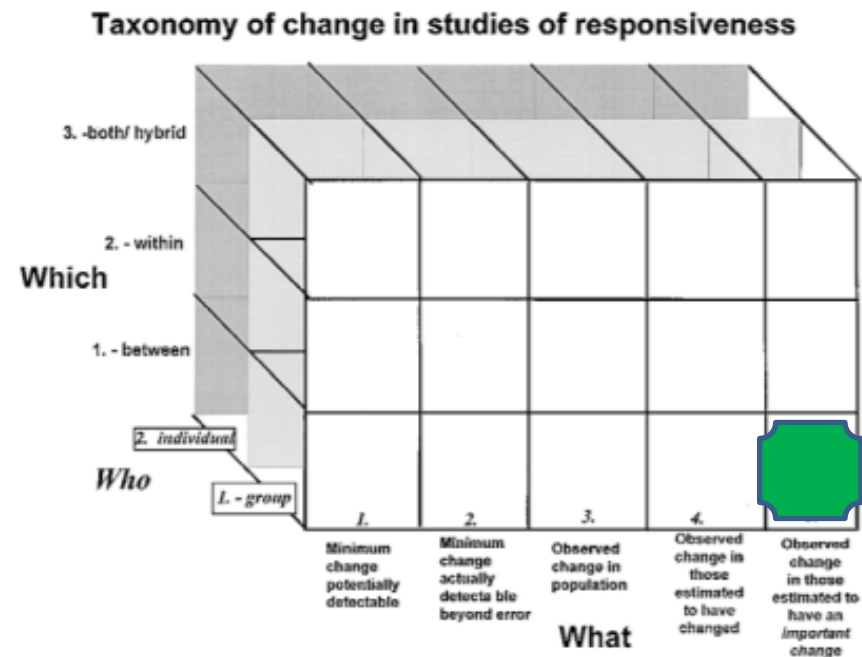


Fig. 5. Depiction of the recommended taxonomy for responsiveness. Each cell in this three-dimensional structure (as defined by the axes of “which, who and what” of the change being quantified), represents a category or type of responsiveness. A given study usually is looking at one kind of change, and therefore one category of responsiveness. A given measure may have information in terms of its responsiveness to one category of change, but this may not mean it is responsive to another.

MCID →

JAMA Guide to Statistics and Methods

**Minimal Clinically Important Difference**  
Defining What Really Matters to Patients

Anna E. McGlothlin, PhD; Roger J. Lewis, MD, PhD



# Example

## Minimal Clinically Important Rehabilitation Effects in Patients with Osteoarthritis of the Lower Extremities

FELIX ANGST, ANDRÉ AESCHLIMANN, BEAT A. MICHEL, and GEROLD STUCKI

**ABSTRACT.** *Objective.* To estimate minimal clinically important differences (MCID) of effects measured by the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) in patients with osteoarthritis (OA) of the lower extremities undergoing a comprehensive inpatient rehabilitation intervention.

*Methods.* A prospective cohort study assessed patients' health by the WOMAC at baseline (entry into the clinic) and at the 3 month followup, and by a transition questionnaire asking about the change of "health in general related to the OA joint" during that time period. The WOMAC section score differences between the "equal" group and the "slightly better" and "slightly worse" groups resulted in the MCID for improvement and for worsening.

*Results.* In total 192 patients were followed up. The MCID for improvement ranged from 0.80 to 1.01 points on the continuous WOMAC numerical rating scale from 0 to 10, reflecting changes of 17 to 22% of baseline scores. The MCID for worsening conditions ranged from 0.29 (6%) to 1.03 points (22%). In the transition reply subjectively unchanged patients reported a "pessimistic bias" of 0.35 to 0.51 points, except for the stiffness section. Both MCID and pessimistic bias showed regression to the mean and baseline dependency.

*Conclusion.* The assessment of MCID using the transition method is a heuristic and valid strategy to detect particular rehabilitation effects in patients with OA of the lower extremities with the use of the WOMAC, and it is worth implementing. The size of the MCID and of the systematic bias is comparable to that assessed by other methods and in other therapeutic settings. (J Rheumatol 2002;29:131–8)

*Key Indexing Terms:*

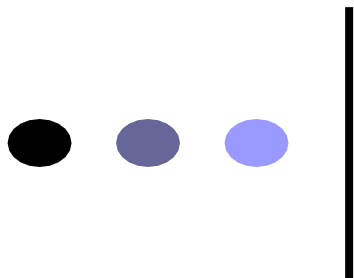
OSTEOARTHRITIS

REHABILITATION

BIOMETRY

WOMAC

SF-36



# Annex



# Hypothesis testing

## Two-sample t-test

Aim: Evaluation of the difference between mean values of two populations

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Assumptions:

- the two samples have normal distribution  $\propto N(M_i, s_i^2)$  and have similar variance ( $s_1^2 \approx s_2^2$ );
- the two samples are independent (unlike the paired t-test)

Degrees of freedom:  $df = n_1 + n_2 - 2$

$$t^* = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Procedure:

- The level of significance is chosen (e.g.,  $\alpha = 0.05$ )
- Compute  $t^*$
- Compare  $t^*$  with the probability distribution table: the null hypothesis is accepted or rejected depending on whether  $t^*$  is lower or higher than  $t_{df, \alpha}$ , respectively



# Hypothesis testing

## Paired two-sample t-test

Aim: Evaluation of the difference between mean values of two *paired* populations

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Assumptions:

- the two samples have normal distribution and are paired ( $n$  pairs)

Degrees of freedom:  $df = n - 1$

$$t^* = \frac{M_1 - M_2}{s_{diff} / \sqrt{n}}$$

where  $s_{diff}$  is the standard deviation of the ordered differences between the two samples

Procedure:

- The level of significance is chosen (e.g.,  $\alpha = 0.05$ )
- Compute  $t^*$
- Compare  $t^*$  with the probability distribution table: the null hypothesis is accepted or rejected depending on whether  $t^*$  is lower or higher than  $t_{df, \alpha}$ , respectively



# Main psychometric properties

*Source: Shirley Ryan Abilitylab*

<https://www.sralab.org/rehabilitation-measures>

## Standard Error of Measurement (SEM)

The Standard Error of Measurement (SEM) is a reliability measure that assesses response stability. The SEM estimates the standard error in a set of repeated scores.

**Clinical Bottom Line:** The SEM is the amount of error that you can consider as measurement error.

*In the Rehabilitation Measures Database, the SEM was frequently pulled directly from peer reviewed journal articles. However, whenever the statistics were available in the published articles, the following equation was utilized to calculate the SEM:*

*$SEM = \text{Standard Deviation from the 1st test} \times (\text{square root of } (1-ICC))$*

SEM is a measurement error in the units used in the measurement.



# Main psychometric properties

Source: Shirley Ryan Abilitylab

<https://www.sralab.org/rehabilitation-measures>

<b>Minimal Detectable Change (MDC)</b>	<p>A statistical estimate of the smallest amount of change that can be detected by a measure that corresponds to a noticeable change in ability.</p> <p><b>Clinical Bottom Line:</b> The MDC is the minimum amount of change in a patient's score that ensures the change isn't the result of measurement error.</p> <p><i>In the Rehabilitation Measures Database, the MDC was frequently pulled directly from peer reviewed journal articles. However, whenever the statistics were available in the published articles, the following equation was utilized to calculate the MDC:</i></p> $MDC = 1.96 \times SEM \times \text{square root of } 2$	<p>The MDC is calculated in terms of confidence of predication. For example, MDC95 is based on a 95% confidence interval, while a MDC90 is based on a 90% confidence interval. Anytime a MDC was calculated for the Rehabilitation Measures Database, the MDC95 was used.</p>
--	--	---





# Main psychometric properties

*Source: Shirley Ryan Abilitylab*

<https://www.sralab.org/rehabilitation-measures>

<b>Minimal Clinically Important Difference (MCID)</b>	<p>MCID represents the smallest amount of change in an outcome that might be considered important by the patient or clinician.</p> <p><b><u>Clinical Bottom Line:</u></b> The MCID is a published value of change in an instrument that indicates the minimum amount of change required for your patient to feel a difference in the variable you are measuring.</p>	<p>The MCID is typically quantified in the units used in the measurement.</p>
---	--	---



# Main psychometric properties

*Source: Shirley Ryan Abilitylab*

<https://www.sralab.org/rehabilitation-measures>

<b>Test-retest Reliability</b>	Establishes that an instrument is capable of measuring a variable with consistency.	<b><u>Clinical Bottom Line:</u></b> If you are planning to use an instrument for <b>individual decision-making</b> , it is recommended that you use an instrument with an <b>ICC &gt; 0.9</b> .  If you are planning to use the instrument to measure progress of a large group (as in research), an instrument with an <b>ICC &gt; 0.7</b> is acceptable.
<b>Interrater Reliability</b>	Determines variation between two or more raters who measure the same group of subjects.	<b><u>Excellent Reliability:</u></b> ICC > 0.75 <b><u>Adequate Reliability:</u></b> ICC 0.40 to < 0.74 <b><u>Poor Reliability:</u></b> ICC < 0.40
<b>Intrarater Reliability</b>	Determines stability of data recorded by one individual across two or more trials.	See Interrater Reliability Criteria