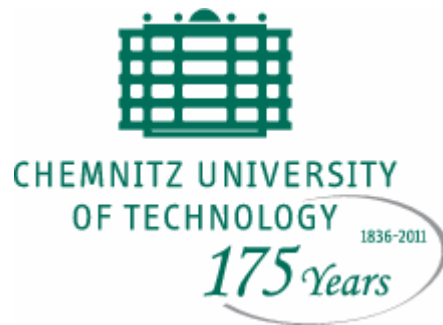


Reliability Assessment in Footwear Biomechanics Research

Christian Maiwald



Overview

1. Why reliability is important
2. Terms and definitions
3. The concept of the **true value** and **measurement error**
4. Methods to quantify reliability
 - a. Relative measures
 - b. Absolute measures
5. Discussion

Theory

Application



Main quality criteria of scientific data collection



Validity

Do the measured variables represent the quantity of interest?

Reliability (Intra-Rater Reliability)

Will repeated measurements under identical boundary conditions deliver the same results?

Objectivity (Inter-Rater Reliability)

Are measurement results independent from the researcher and the boundary conditions of the experiment?



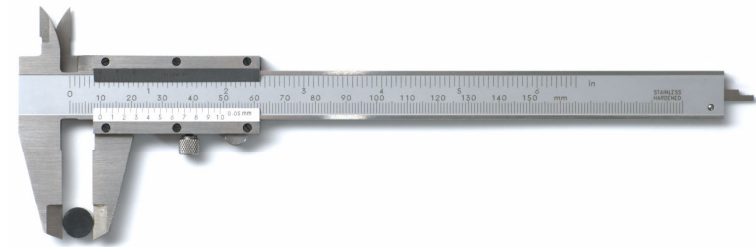
Benefit of reliability information

- Quality of the research protocol and instrumentation
- Comparability of own data, protocols & results to published work
- Interpretation of results
Unreliable measurements vs. small effects
- Rating of relevant effect magnitude
Calculation of sample sizes
- ...



Reliability: „quality of measurements“

- Reliability is a characteristic of the entire measurement process, not the measuring instrument or the thing being measured
- The more reliable a measurement is, the more we may describe it as being „precise“ and „accurate“



Definition of *reliability*



Reliability can be defined as the consistency of measurements, or of an individual's performance, on a test; or 'the absence of measurement error'.^[17]

[Atkinson & Nevill 1998]

measured quantity = true quantity + measurement error

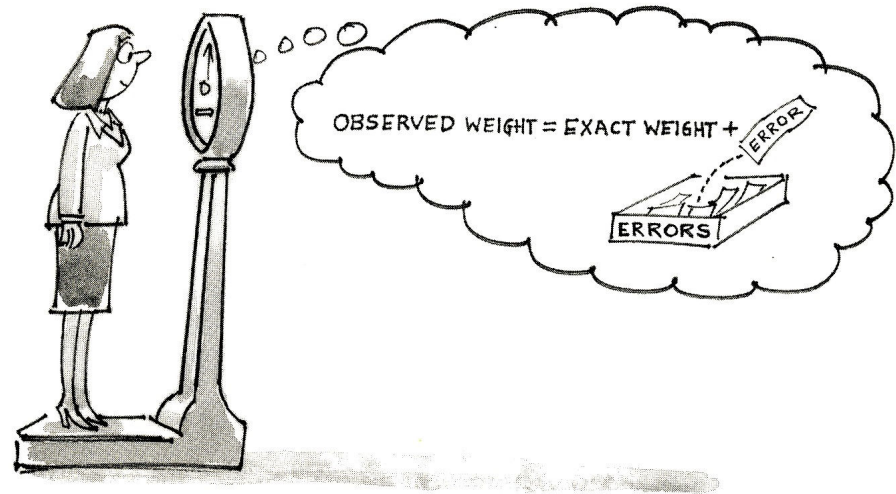
(„measurement“: any form of collecting quantitative data)

Consistency vs. Agreement!



A model for measurement error

- Measurement error is thought to occur by chance
- Box model
 - Measurement process: draw with replacement from a box of tickets (error box)
 - Error box mean = 0
 - Error box SD: unknown



[Freedman et al. 2007]

The **SD of the error box** yields the likely size of random error we can expect for each measurement. It can be determined from the outcome of repeated measurements under **identical boundary conditions**.

Possible sources of measurement error

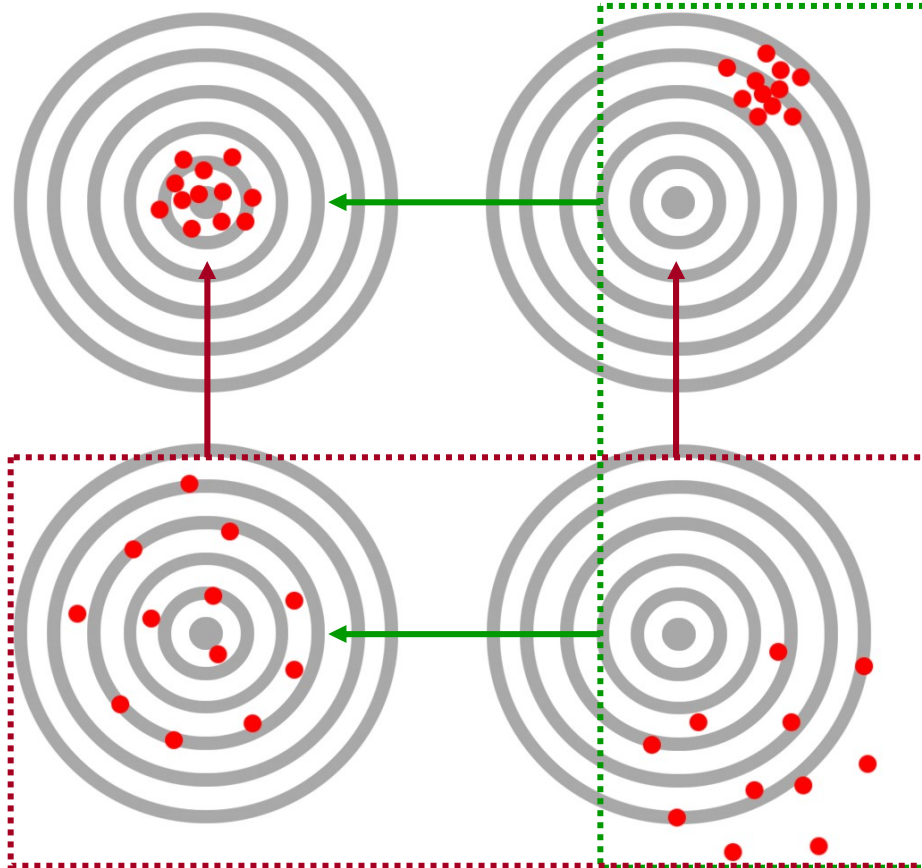
- **Instrumentation**
 - Technical deficits, wear, drift
 - repercussion on the subject
 - Modelling and data processing
(e. g. soft tissue artifacts in kinematic data)
- **Subjects („target“)**
 - Physiological variability in performing tasks
 - Adaptation to measurement situation
(e. g. routine, fatigue)
- **Researcher („rater“, „judge“)**
 - Operation of instruments
 - Instruction of subjects
(e. g. when using questionnaires or conducting laboratory experiments)
 - Subjective ratings / biased raters / skill level and experience
(e. g. palpation, clinical examination)



Random & systematic components of measurement error: the target analogy

no systematic error
good *accuracy*
unbiased

systematic error,
lack of *accuracy*
biased



no random error,
good *precision*

Instrument calibration
Operator training

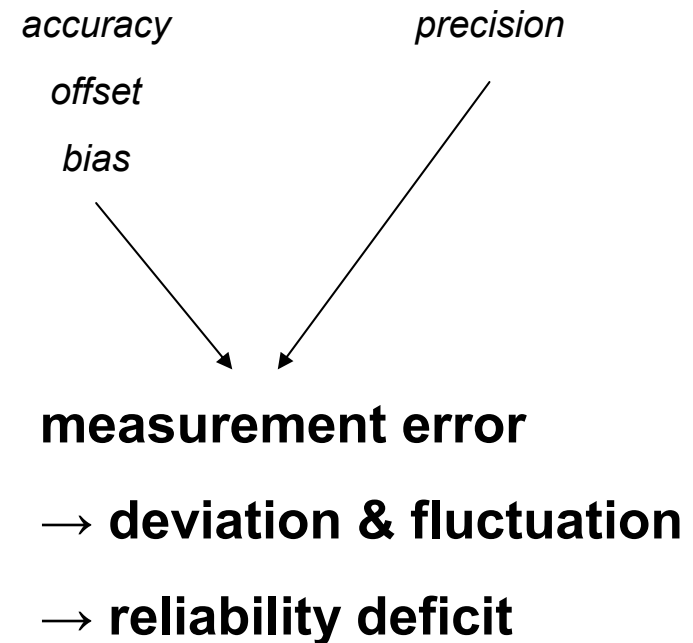
random error,
lack of *precision*

„defective“
instrumentation or
protocol?
Unexperienced
researcher?
Identical boundary
conditions for
every shot?



Random & systematic components of measurement error

Measured quantity = true quantity + (systematic error + random error)



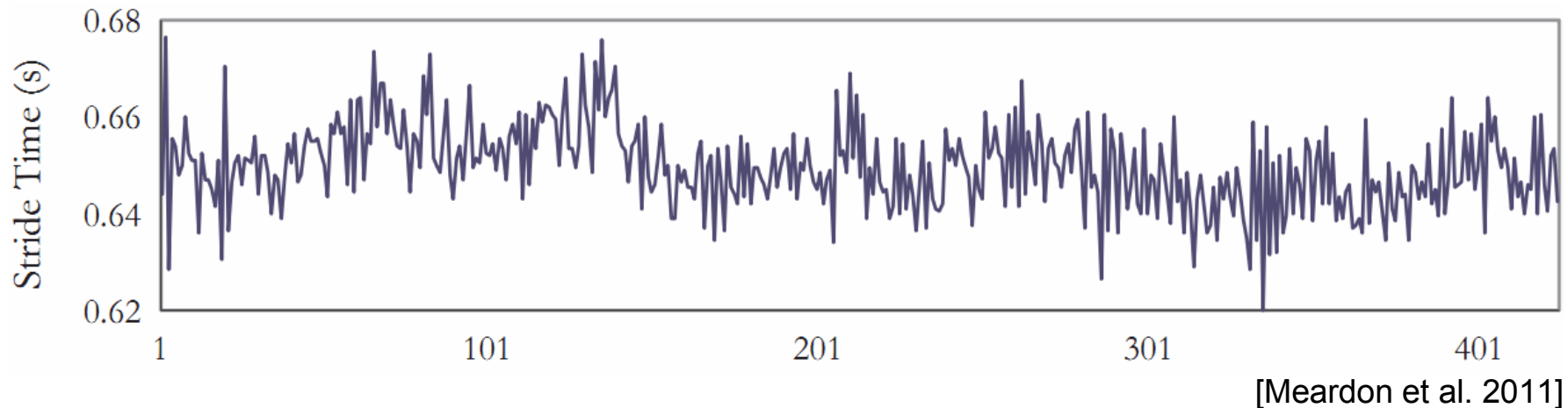
Random & systematic components of measurement error

- Multiple ways to compensate for systematic error!
 - Calibration of the instrument
 - Modification of the research protocol
(e. g. control for fatigue, include accommodation sessions...)
 - Training of research staff
 - ...
- No simple way to eliminate random error
 - Improve instrumentation?
 - Training of research staff?
 - Are we aiming at a „moving target“?

Is the variable characteristic of human performance (physiological variability) the main source of random error?



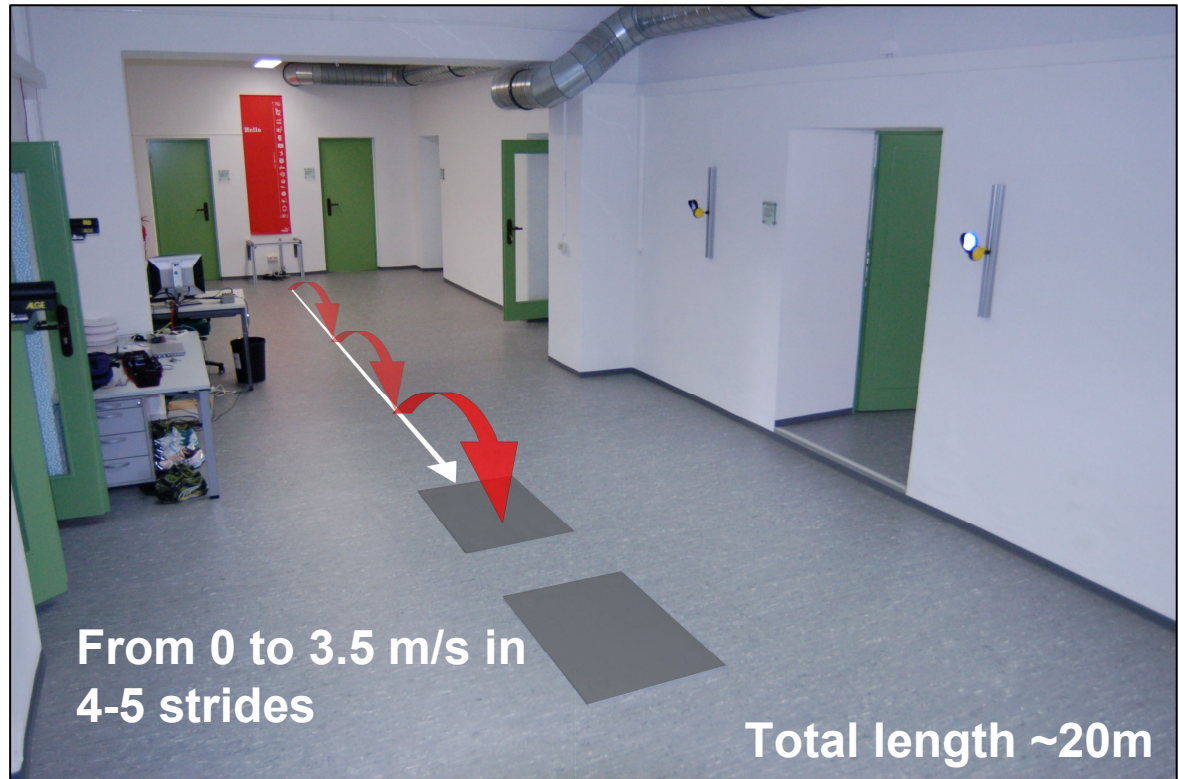
Physiological variability: Running



- Step to step fluctuations in any measured variable!
- Is this (random) error or true value fluctuation?
Theory presumes a constant “true value” underlying this data. Thus, fluctuation will be considered as “measurement error”!
- Mean value serves as *best estimate* of a constant true value of a subject

Physiological variability & running under laboratory conditions

- Limited spacial dimensions
- Usually 5-10 repeated trials
- Running: one or two steps per trial with data records
- Total: 10-20 steps



Common procedures presume the mean value of the recorded laboratory trials to be a *valid estimator* of the true value of a subject!

Visit Poster #7 tomorrow for more information!

Preliminary summary

- Reliability is a prerequisite to achieve valid measurements
- Classical measurement theory presumes an underlying **true value** of a subject, that is to be captured
- **Measurement error is a...**
 - ...**lack of consistency** of repeated measurements
 - ...**lack of agreement** between the true quantity and the result of the measurement process
- Measurement error can be separated into **systematic** and **random** components



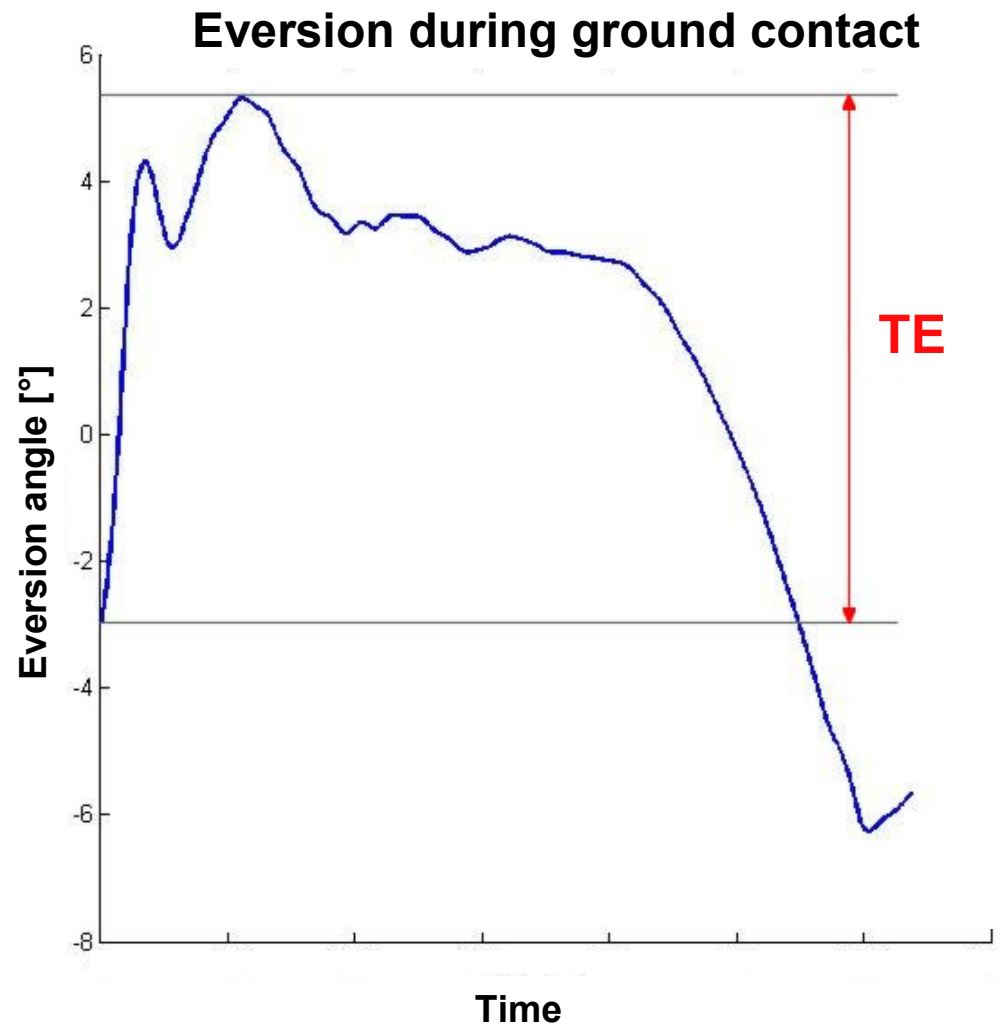
How much
measurement error is present
in our data?

Methods to quantify reliability



Sample dataset of goniometer records

- 24 runners
- **Total Eversion TE [°] measured via goniometer**
- 5 valid trials per session at 3.5 m/s
- 2 sessions
~ 4 weeks apart
- No intervention
inbetween sessions
- Same neutral
running shoe



Sample dataset of goniometer records

- 24 runners
- **Total Eversion TE [°] measured via goniometer**
- 5 valid trials per session at 3.5 m/s
- 2 sessions
~ 4 weeks apart
- No intervention inbetween sessions
- Same neutral running shoe
- 3 data subsets:
Session1
Session2
Means

Subject	Session 1					Mean S1	Session 2					Mean S2
	T1	T2	T3	T4	T5		T6	T7	T8	T9	T10	
S01	13.4	11.3	10.7	11.7	14.3	12.3	13.1	11.8	10.1	12.4	14.5	12.4
S02	13.6	16.1	18.3	19.0	19.1	17.2	20.9	21.2	19.6	19.8	19.6	20.2
S03	11.1	9.3	10.3	9.5	8.7	9.8	8.8	11.5	11.2	9.7	9.7	10.2
S04	11.9	14.5	12.6	13.2	12.2	12.9	14.1	14.0	15.0	15.0	11.0	13.8
S05	12.9	16.7	14.0	15.4	13.7	14.6	15.6	13.7	13.9	14.3	14.7	14.4
S06	11.5	13.9	12.3	9.6	12.7	12.0	13.0	12.7	13.1	13.0	11.5	12.7
S07	12.9	14.3	13.1	12.6	12.1	13.0	17.3	13.2	13.4	13.8	10.7	13.7
S08	11.2	15.1	12.9	12.7	13.6	13.1	12.5	12.7	11.9	10.3	12.1	11.9
S09	12.2	10.3	11.1	9.6	12.2	11.1	11.1	13.4	11.8	13.5	12.8	12.5
S10	10.7	14.5	13.1	10.7	9.8	11.8	11.5	10.8	10.9	11.9	10.5	11.1
S11	14.4	14.2	15.4	15.0	14.1	14.6	14.3	15.3	15.2	15.0	16.0	15.2
S12	6.0	7.9	10.0	6.7	8.7	7.9	6.9	6.7	6.9	8.3	7.4	7.2
S13	12.9	13.3	11.9	16.7	13.4	13.6	15.5	15.5	18.0	17.1	15.6	16.4
S14	16.5	17.2	16.9	16.8	17.3	16.9	14.3	13.6	11.4	13.1	13.1	13.1
S15	13.3	13.9	13.8	15.0	14.0	14.0	15.8	15.1	13.0	14.2	16.0	14.8
S16	12.4	10.7	11.1	11.8	12.4	11.7	12.5	11.4	14.2	12.9	12.4	12.7
S17	20.4	21.5	22.7	19.3	22.3	21.2	22.9	23.2	24.3	27.5	23.2	24.2
S18	17.3	15.4	15.9	16.3	17.0	16.4	14.2	16.0	15.3	15.2	16.2	15.4
S19	17.5	14.2	17.3	12.9	14.1	15.2	12.6	12.1	14.5	12.7	14.3	13.2
S20	10.6	10.2	10.3	11.4	8.2	10.1	11.8	10.4	12.1	10.7	9.6	10.9
S21	6.3	5.4	6.5	4.7	5.4	5.7	6.1	5.0	6.6	6.7	5.7	6.0
S22	11.9	10.2	11.3	11.9	9.3	10.9	9.9	10.1	8.7	10.1	9.2	9.6
S23	12.3	13.2	12.7	10.6	14.0	12.5	12.5	11.2	14.3	12.8	13.7	12.9
S24	13.4	16.8	12.9	14.9	14.9	14.6	13.5	13.4	11.7	15.1	14.3	13.6

Measures of reliability

Relative reliability is the degree to which individuals maintain their position in a sample with repeated measurements. This type of reliability is usually assessed with some type of correlation coefficient. Absolute reliability is the degree to which repeated measurements vary for individuals. This type of reliability is expressed either in the actual units of measurement or as a proportion of the measured values (dimensionless ratio).

[Atkinson & Nevill 1998]

Popular measures of **relative** reliability:

- Intraclass Correlation Coefficient (ICC)
- Pearson's r (test-retest)

Popular measures of **absolute** reliability:

- Limits of Agreement
- Root Mean Square Error (similar: SEM, TEM)
- Coefficient of Variation (CV)

Box model!

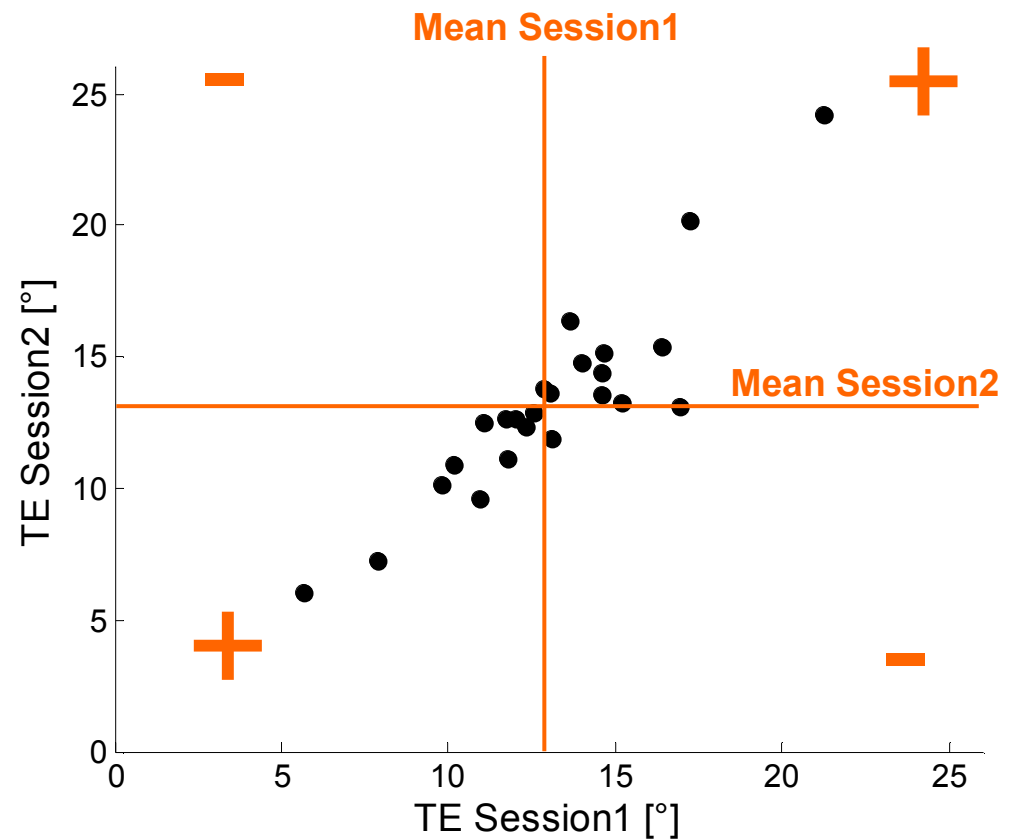


Measures of relative reliability

Test-retest correlation (Pearson's r)

- Presumes linear relationship between repeated measurements
- Calculated as mean of standard unit products

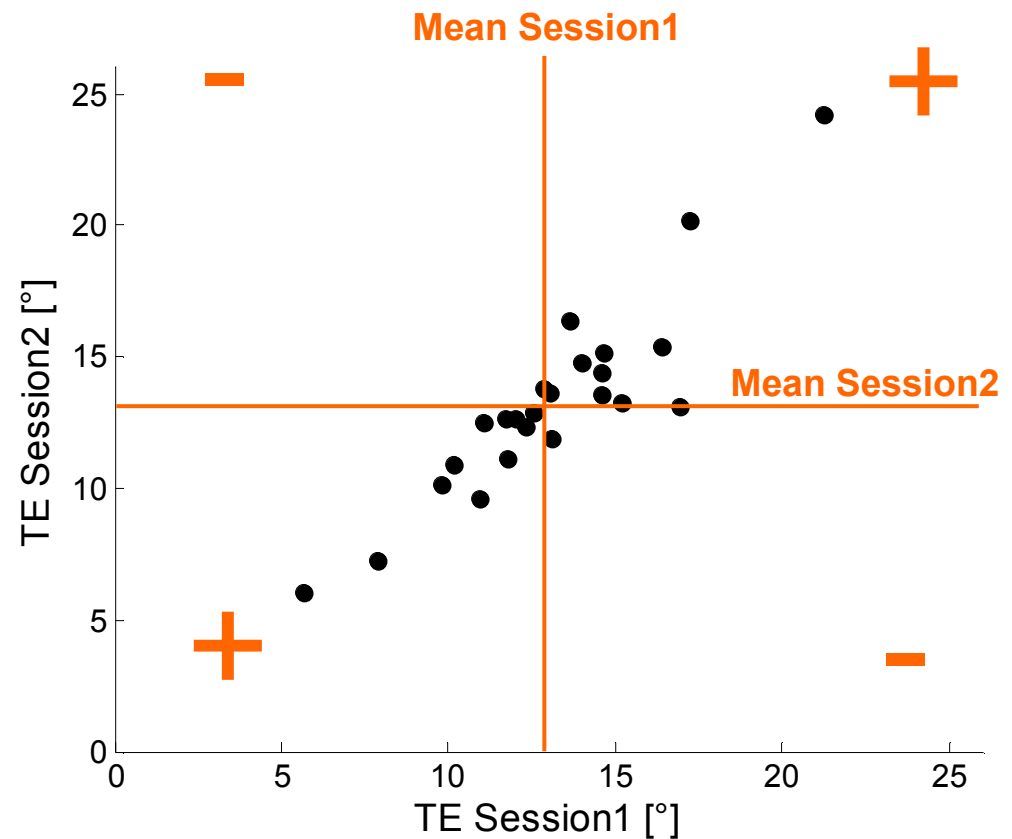
**Green dataset:
 $r = 0.91$**



Measures of relative reliability

Test-retest correlation (Pearson's r)

- Simple and easy to accomplish
- Measure of linear relationship (and consistence), not of agreement!
Unable to detect systematic offset!
- Interpretation according to „arbitrary“ thresholds



Measures of relative reliability

Intraclass Correlation Coefficient

- Applies to two or more repetitions
(blue and green datasets)
- Ratio of variances (between/within subjects)
- Presumes repeated measurements to be exchangeable
- At least 6 different statistical models and calculations for different applications
(e.g. study designs)

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

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

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

$$ICC(1, k) = (BMS - WMS)/BMS$$

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

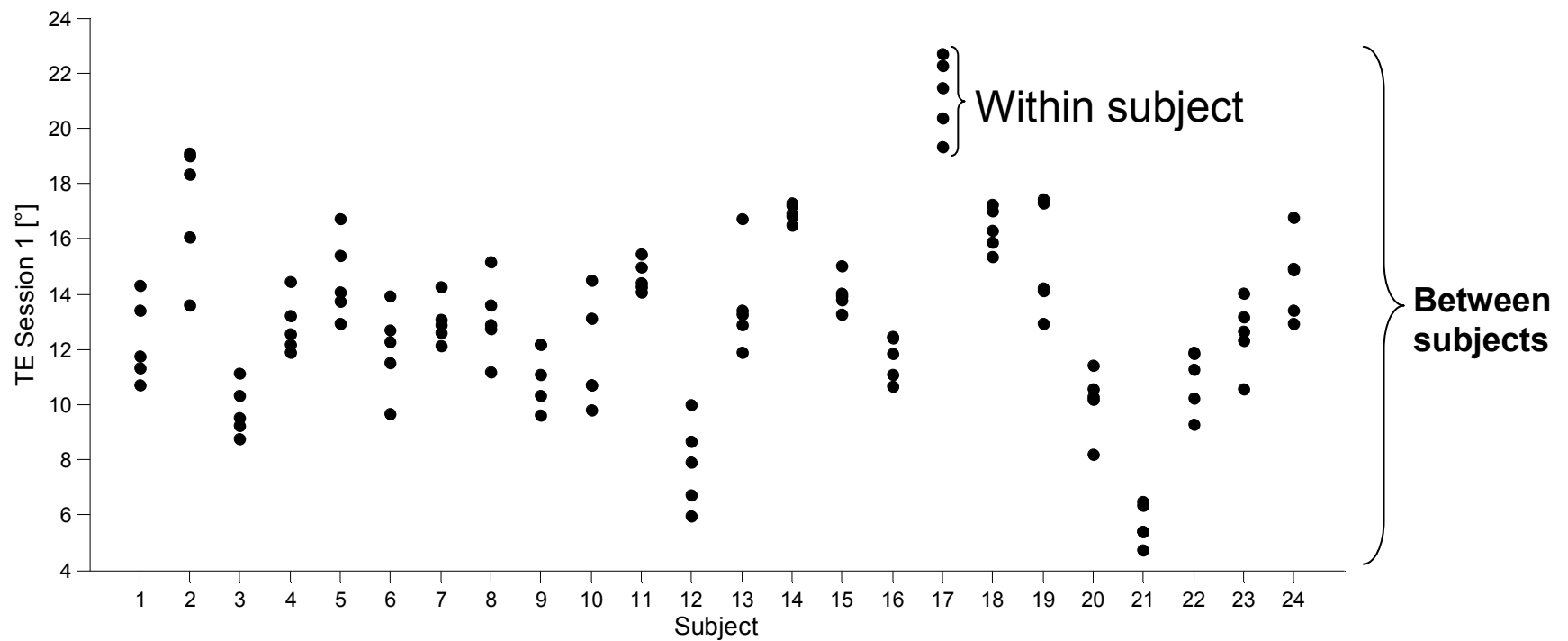
$$ICC(3, k) = (BMS - EMS)/BMS$$

[Shrout & Fleiss 1979]



Measures of relative reliability

Intraclass Correlation Coefficient



Measures of relative reliability

Intraclass Correlation Coefficient:

- Able to detect systematic offset
(Some coefficients, not all of them!)
- Applicable to many study designs
- Presumes substantial between subjects variance!
- Interpretation according to „arbitrary“ thresholds

ICC type	Session1	Session2	Session Means
ICC(1,1)	0.86	0.91	0.90
ICC(1,k)	0.96	0.98	0.95
ICC(2,1)	0.86	0.91	0.90
ICC(2,k)	0.96	0.98	0.95
ICC(3,1)	0.84	0.91	0.90
ICC(3,k)	0.96	0.98	0.95



Measures of absolute reliability: Limits of Agreement (LoA)

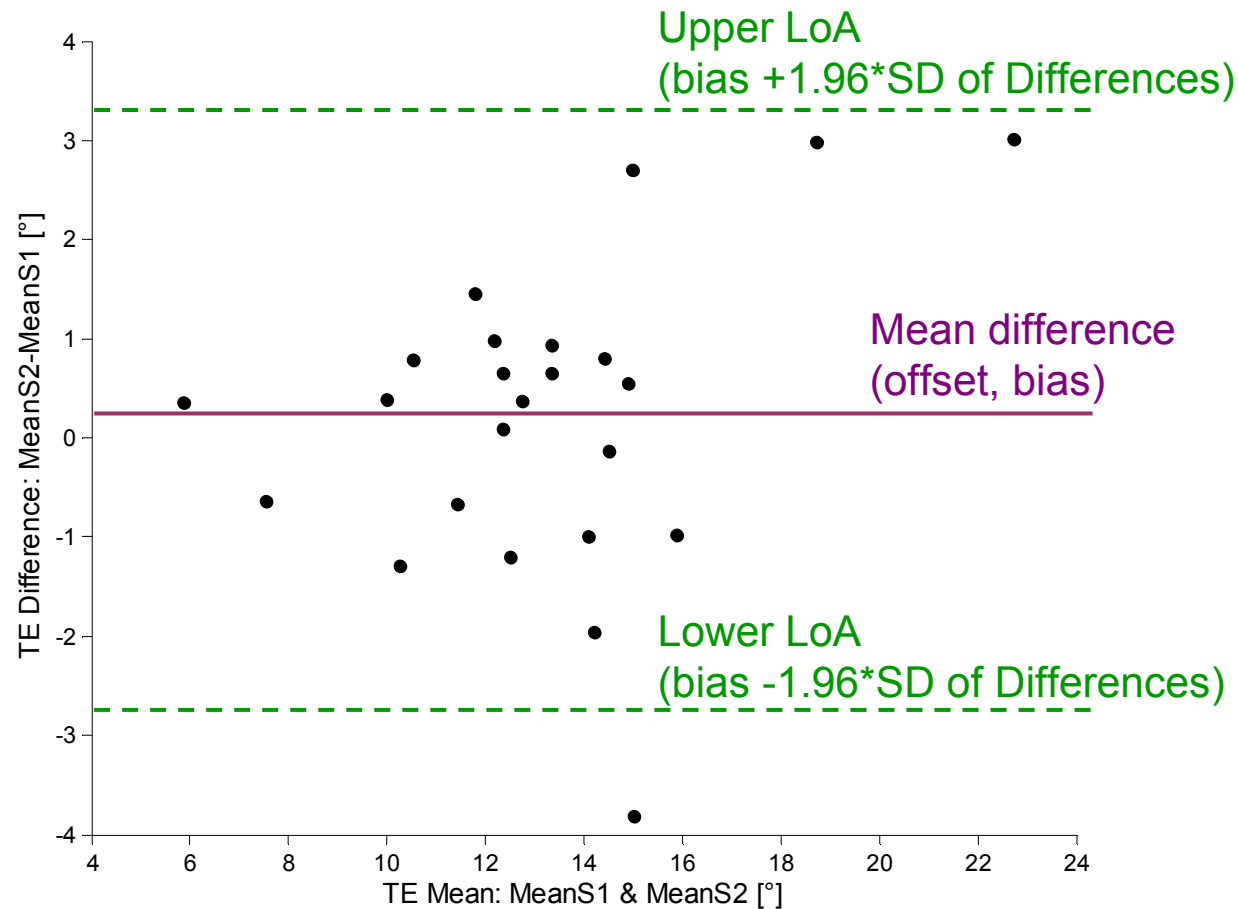
2 measurements

Subject	Session 1						Session 2					
	T1	T2	T3	T4	T5	Mean S1	T6	T7	T8	T9	T10	Mean S2
S01	13.4	11.3	10.7	11.7	14.3	12.3	13.1	11.8	10.1	12.4	14.5	12.4
S02	13.6	16.1	18.3	19.0	19.1	17.2	20.9	21.2	19.6	19.8	19.6	20.2
S03	11.1	9.3	10.3	9.5	8.7	9.8	8.8	11.5	11.2	9.7	9.7	10.2
S04	11.9	14.5	12.6	13.2	12.2	12.9	14.1	14.0	15.0	15.0	11.0	13.8
S05	12.9	16.7	14.0	15.4	13.7	14.6	15.6	13.7	13.9	14.3	14.7	14.4
S06	11.5	13.9	12.3	9.6	12.7	12.0	13.0	12.7	13.1	13.0	11.5	12.7
S07	12.9	14.3	13.1	12.6	12.1	13.0	17.3	13.2	13.4	13.8	10.7	13.7
S08	11.2	15.1	12.9	12.7	13.6	13.1	12.5	12.7	11.9	10.3	12.1	11.9
S09	12.2	10.3	11.1	9.6	12.2	11.1	11.1	13.4	11.8	13.5	12.8	12.5
S10	10.7	14.5	13.1	10.7	9.8	11.8	11.5	10.8	10.9	11.9	10.5	11.1
S11	14.4	14.2	15.4	15.0	14.1	14.6	14.3	15.3	15.2	15.0	16.0	15.2
S12	6.0	7.9	10.0	6.7	8.7	7.9	6.9	6.7	6.9	8.3	7.4	7.2
S13	12.9	13.3	11.9	16.7	13.4	13.6	15.5	15.5	18.0	17.1	15.6	16.4
S14	16.5	17.2	16.9	16.8	17.3	16.9	14.3	13.6	11.4	13.1	13.1	13.1
S15	13.3	13.9	13.8	15.0	14.0	14.0	15.8	15.1	13.0	14.2	16.0	14.8
S16	12.4	10.7	11.1	11.8	12.4	11.7	12.5	11.4	14.2	12.9	12.4	12.7
S17	20.4	21.5	22.7	19.3	22.3	21.2	22.9	23.2	24.3	27.5	23.2	24.2
S18	17.3	15.4	15.9	16.3	17.0	16.4	14.2	16.0	15.3	15.2	16.2	15.4
S19	17.5	14.2	17.3	12.9	14.1	15.2	12.6	12.1	14.5	12.7	14.3	13.2
S20	10.6	10.2	10.3	11.4	8.2	10.1	11.8	10.4	12.1	10.7	9.6	10.9
S21	6.3	5.4	6.5	4.7	5.4	5.7	6.1	5.0	6.6	6.7	5.7	6.0
S22	11.9	10.2	11.3	11.9	9.3	10.9	9.9	10.1	8.7	10.1	9.2	9.6
S23	12.3	13.2	12.7	10.6	14.0	12.5	12.5	11.2	14.3	12.8	13.7	12.9
S24	13.4	16.8	12.9	14.9	14.9	14.6	13.5	13.4	11.7	15.1	14.3	13.6



Measures of absolute reliability: Limits of Agreement (LoA)

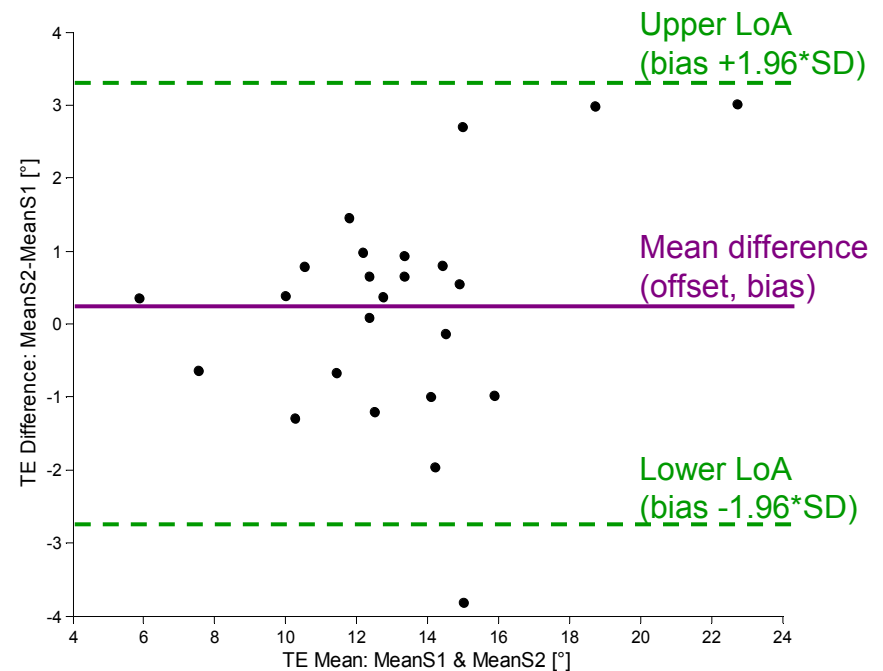
2 measurements



We can expect 95% of the differences to lie in this interval.
Sample data: $[0.21^\circ \pm 3.04^\circ]$

Interpretation of the LoA

- Is there a relationship between variable magnitude and difference of measurements?
(→ heteroscedasticity?)
- How do the differences distribute?
Random? 95% within LoA?
- **What's the magnitude of the LoA? Can we accept the obtained LoA as an interval for random fluctuations?**



Measures of absolute reliability: Root Mean Square Error (RMSE)

2 or more measurements

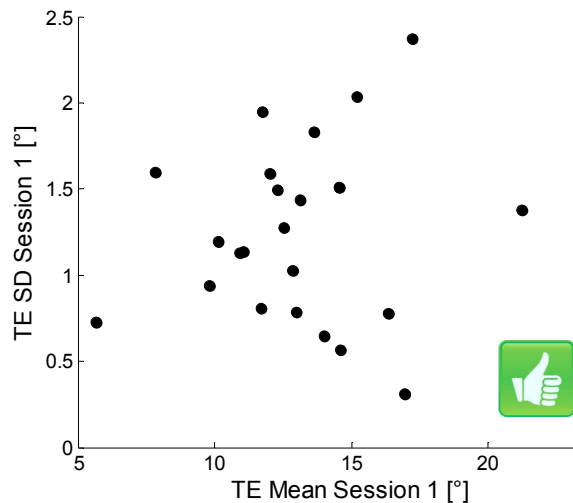
Subject	Session 1					Mean S1	Session 2					Mean S2
	T1	T2	T3	T4	T5		T6	T7	T8	T9	T10	
S01	13.4	11.3	10.7	11.7	14.3	12.3	13.1	11.8	10.1	12.4	14.5	12.4
S02	13.6	16.1	18.3	19.0	19.1	17.2	20.9	21.2	19.6	19.8	19.6	20.2
S03	11.1	9.3	10.3	9.5	8.7	9.8	8.8	11.5	11.2	9.7	9.7	10.2
S04	11.9	14.5	12.6	13.2	12.2	12.9	14.1	14.0	15.0	15.0	11.0	13.8
S05	12.9	16.7	14.0	15.4	13.7	14.6	15.6	13.7	13.9	14.3	14.7	14.4
S06	11.5	13.9	12.3	9.6	12.7	12.0	13.0	12.7	13.1	13.0	11.5	12.7
S07	12.9	14.3	13.1	12.6	12.1	13.0	17.3	13.2	13.4	13.8	10.7	13.7
S08	11.2	15.1	12.9	12.7	13.6	13.1	12.5	12.7	11.9	10.3	12.1	11.9
S09	12.2	10.3	11.1	9.6	12.2	11.1	11.1	13.4	11.8	13.5	12.8	12.5
S10	10.7	14.5	13.1	10.7	9.8	11.8	11.5	10.8	10.9	11.9	10.5	11.1
S11	14.4	14.2	15.4	15.0	14.1	14.6	14.3	15.3	15.2	15.0	16.0	15.2
S12	6.0	7.9	10.0	6.7	8.7	7.9	6.9	6.7	6.9	8.3	7.4	7.2
S13	12.9	13.3	11.9	16.7	13.4	13.6	15.5	15.5	18.0	17.1	15.6	16.4
S14	16.5	17.2	16.9	16.8	17.3	16.9	14.3	13.6	11.4	13.1	13.1	13.1
S15	13.3	13.9	13.8	15.0	14.0	14.0	15.8	15.1	13.0	14.2	16.0	14.8
S16	12.4	10.7	11.1	11.8	12.4	11.7	12.5	11.4	14.2	12.9	12.4	12.7
S17	20.4	21.5	22.7	19.3	22.3	21.2	22.9	23.2	24.3	27.5	23.2	24.2
S18	17.3	15.4	15.9	16.3	17.0	16.4	14.2	16.0	15.3	15.2	16.2	15.4
S19	17.5	14.2	17.3	12.9	14.1	15.2	12.6	12.1	14.5	12.7	14.3	13.2
S20	10.6	10.2	10.3	11.4	8.2	10.1	11.8	10.4	12.1	10.7	9.6	10.9
S21	6.3	5.4	6.5	4.7	5.4	5.7	6.1	5.0	6.6	6.7	5.7	6.0
S22	11.9	10.2	11.3	11.9	9.3	10.9	9.9	10.1	8.7	10.1	9.2	9.6
S23	12.3	13.2	12.7	10.6	14.0	12.5	12.5	11.2	14.3	12.8	13.7	12.9
S24	13.4	16.8	12.9	14.9	14.9	14.6	13.5	13.4	11.7	15.1	14.3	13.6



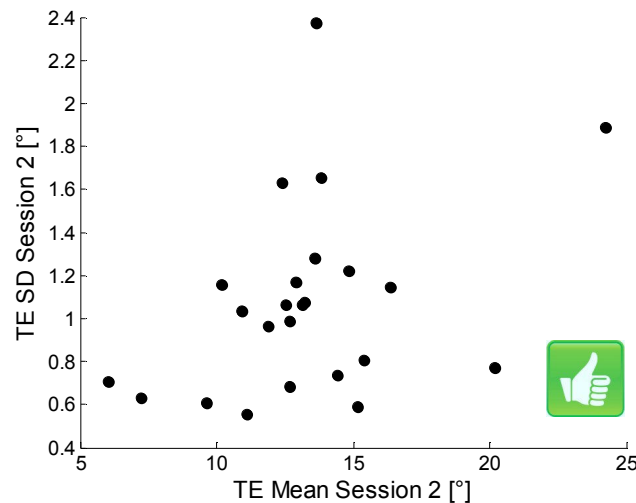
Presumptions for RMSE

Variable magnitude should be unrelated to within subject variability!

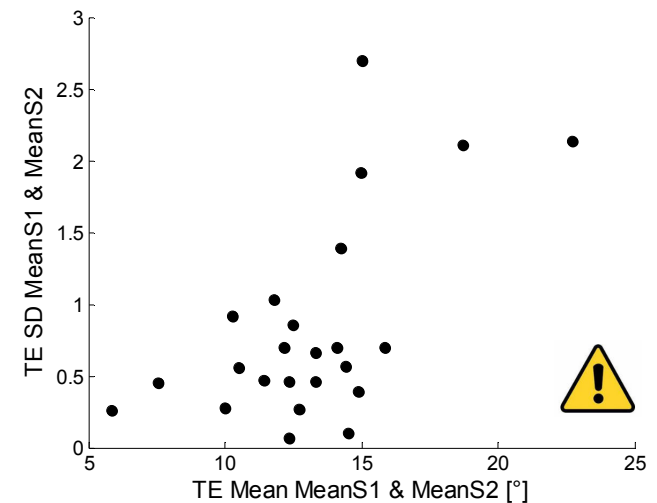
→ check via scatterplot / (rank) correlation!



$r = 0.12 / \tau = 0.04$



$r = 0.37 / \tau = 0.25$



$r = 0.61 / \tau = 0.36$



Measures of absolute reliability: Root Mean Square Error (RMSE)

RMSE is calculated as the square root of the residual error mean square of a *two-way* ANOVA:

	Source	SS	df	MS	F	p
SESSION1	Subjects	1188.2	23	51.7	28.1	0
	Trials	5.3	4	1.3	0.7	0.58
	Error	169.0	92	1.8		
	Total	1362.5	119			
SESSION2	Subjects	1581.6	23	68.8	50.5	0
	Trials	3.9	4	1.0	0.7	0.58
	Error	125.3	92	1.4		
	Total	1710.8	119			
MEANS	Subjects	526.3	23	22.9	19.0	0
	Sessions	0.5	1	0.5	0.4	0.51
	Error	27.7	23	1.2		
	Total	554.5	47			

ANOVA indicates the presence of **systematic error** due to trials / sessions

Measures of absolute reliability: Root Mean Square Error (RMSE)

RMSE reflects the random error component [Massé et al. 1997]:

	Source	SS	df	MS	F	p	RMSE [°]
SESSION1	Subjects	1188.2	23	51.7	28.1	0	
	Trials	5.3	4	1.3	0.7	0.58	
	Error	169.0	92	1.8			$\sqrt{1.8} \approx 1.35$
	Total	1362.5	119				
SESSION2	Subjects	1581.6	23	68.8	50.5	0	
	Trials	3.9	4	1.0	0.7	0.58	
	Error	125.3	92	1.4			$\sqrt{1.4} \approx 1.18$
	Total	1710.8	119				
MEANS	Subjects	526.3	23	22.9	19.0	0	
	Sessions	0.5	1	0.5	0.4	0.51	
	Error	27.7	23	1.2			$\sqrt{1.2} \approx 1.10$
	Total	554.5	47				

Interpretation of RMSE

„The difference between a subject's measurement and the true value would be expected to be less than $1.96 \times RMSE$ for 95% of observations.“ [Bland & Altman, 1996]

Interval for the true quantity of a subject: [measured quantity $\pm 1.96 \times RMSE$]

→ blue datasets!

Sample data: $[x \pm 2.64^\circ]$ and $[x \pm 2.31^\circ]$

Interpretation of RMSE

„Another useful way of presenting measurement error is sometimes called the *repeatability*, which is $\sqrt{2} \times 1.96 \times RMSE$ or $2.77 \times RMSE$.

The difference between two measurements for the same subject is expected to be less than $2.77 \times RMSE$ for 95% of observations.“

[Bland & Altman, 1996]

Interval for the true difference of a quantity: [measured difference $\pm 2.77 \times RMSE$]

→ green dataset!

Sample data: [difference $\pm 3.04^\circ$] → LoA!



Summary: Relative measures of reliability

- Popular, widespread use, well accepted among researchers
- Reliability „condensed to a single number“
- Comparability is rather difficult
 - Dependent on the range of the data
(between subject variance!)
 - Dependent on the type of coefficient
(e.g. choosing the appropriate ICC is not trivial)
- Conclusion
 - Dimensionless single number, but data magnitudes are „hidden“
 - Arbitrary thresholds for interpretation (what is poor, what is excellent?)



Summary: Absolute measures of reliability

- Slowly growing popularity, necessity to convince reviewers and readers (although the Bland & Altman paper is among the most frequently cited in medical science!)
- Assumption of a box model for measurement error
Box model applies for stationary objects. Does it apply to human performance?
- Implicit rating of obtained results
→ LoA and RMSE are given in the unit of interest!
- Comparability is easy:
 - Independent from sample size or range of data
 - Well documented, simple calculation
- Conclusion
 - Much easier to interpret
 - Random fluctuation ↔ Intervals of „non-relevance“?
→ Calculation of sample size



Discussion

- Are we asking for **consistency** or **agreement**?
→ relative vs. Absolute measures of reliability!
- **The true value** of a subject
 - Is e. g. running an example of repeating steps with an underlying constant and random fluctuation? Does the model actually fit human movement characteristics?
 - How do research protocols reflect this issue?
- Laboratory conditions
 - Identical boundary conditions?
 - Natural variability that allows the estimation of a true value?
- Strategies for handling human performance variability
 - Sample sizes, number of repeated trials?
 - Field testing? Laboratory conditions required for certain applications!
 - How to control for subject adaptation to measurement conditions (e.g. treadmill, isokinetic devices...)
- Consequences
 - How can reliability studies create the most benefit for the research community?



References and recommended literature

G. Atkinson & A. M. Nevill:

*Statistical Methods For Assessing Measurement Error (Reliability)
in Variables Relevant to Sports Medicine*

Sports Med 1998, Vol. 26 (4), p217-238

Shrout, P. & Fleiss, J.

Intraclass correlations: uses in assessing rater reliability

Psychol Bull, 1979, 86, 420-428

Massé, J.; Bland, J. M. & Doyle, J.

Measurement error. A constant within subject standard deviation cannot be assumed a priori.

British Medical Journal 1997, Vol. 314, p147

Bland, J. M. & Altman, D. G.

Measurement error

British Medical Journal 1996, Vol. 313, p744

Bland, J. M. & Altman, D. G.

Measuring Agreement in method comparison studies

Statistical Methods in Medical Research 1999, Vol. 8, p135-160

Freedman, D.; Pisani, R. & Purves, R.

Statistics

4th Edition, New York 2008
ISBN 0-393-92972-8



Thank you for your attention!

Christian Maiwald
Institute of Sport Science
Chemnitz University of Technology
09107 Chemnitz, Germany

christian.maiwald@hsw.tu-chemnitz.de

