"Research methods to critically appraise measurement proprieties in pain measurement"

Raymond Ostelo
Professor of Evidence Based Physiotherapy
EMGO+ Institute

VU medisch centrum
EMGO+ Institute for Health and Care Research

Some backgound

Rehabilitation after lumbar disc surgery (Review)

Oosterhuis T, Costa LOP, Maher CG, de Vet HCW, van Tulder MW, Ostelo RWJG

THE COCHRANE
COLLABORATION®

Some backgound

SPINE Volume 28, Number 16, pp 1757–1765

Behavioral Graded Activity Following First-Time Lumbar Disc Surgery

1-Year Results of a Randomized Clinical Trial

Raymond W. J. G. Ostelo, PT, PhD,*† Henrica C. W. de Vet, PhD,‡
Johan W. S. Vlaeyen, PhD,§ Maria R. Kerckhoffs, PT, MSc,*†|| Willem M. Berfelo, MD PhD,¶
Pieter M. J. C. Wolters, PT,|| and Piet A. van den Brandt, PhD*

SPINE Volume 29, Number 6, pp 615–622

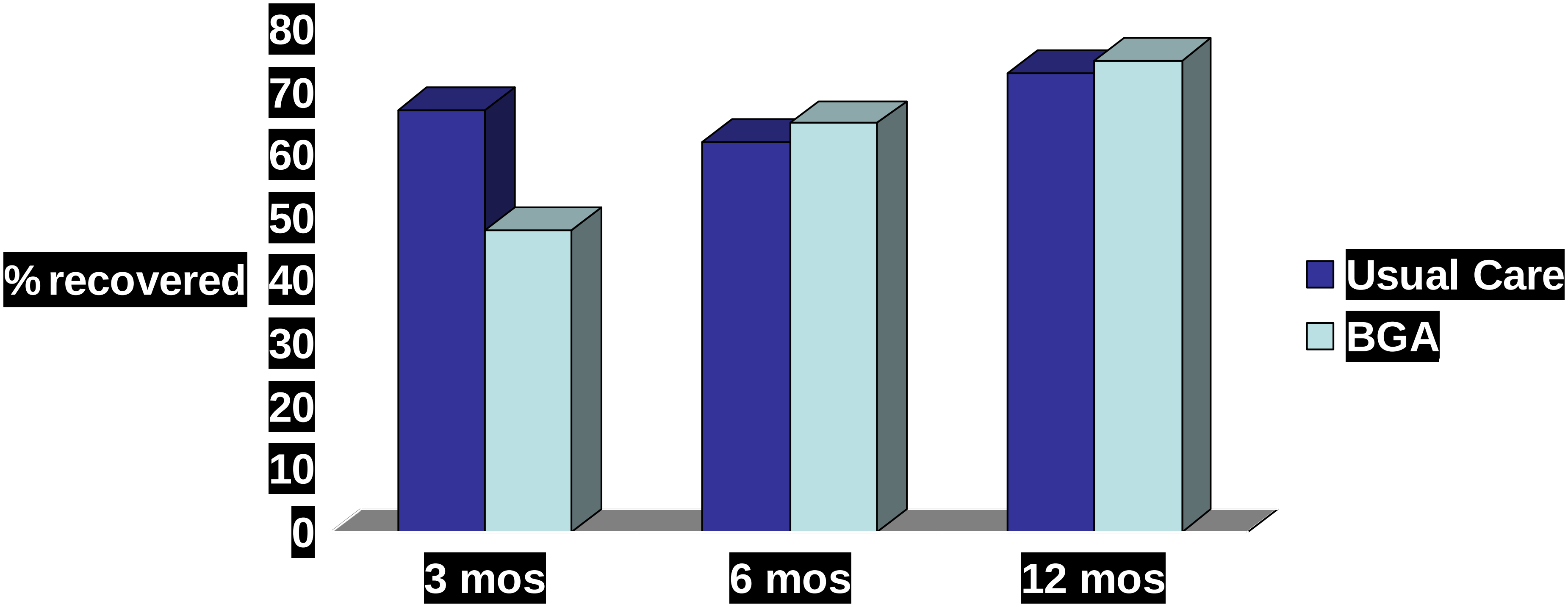
Economic Evaluation of a Behavioral-Graded Activity Program Compared to Physical Therapy for Patients Following Lumbar Disc Surgery

Raymond W. J. G. Ostelo, PhD, PT,*‡ Mariëlle E. J. B. Goossens, PhD,†§
Henrica C. W. de Vet, PhD,‡ and Piet A. van den Brandt, PhD*

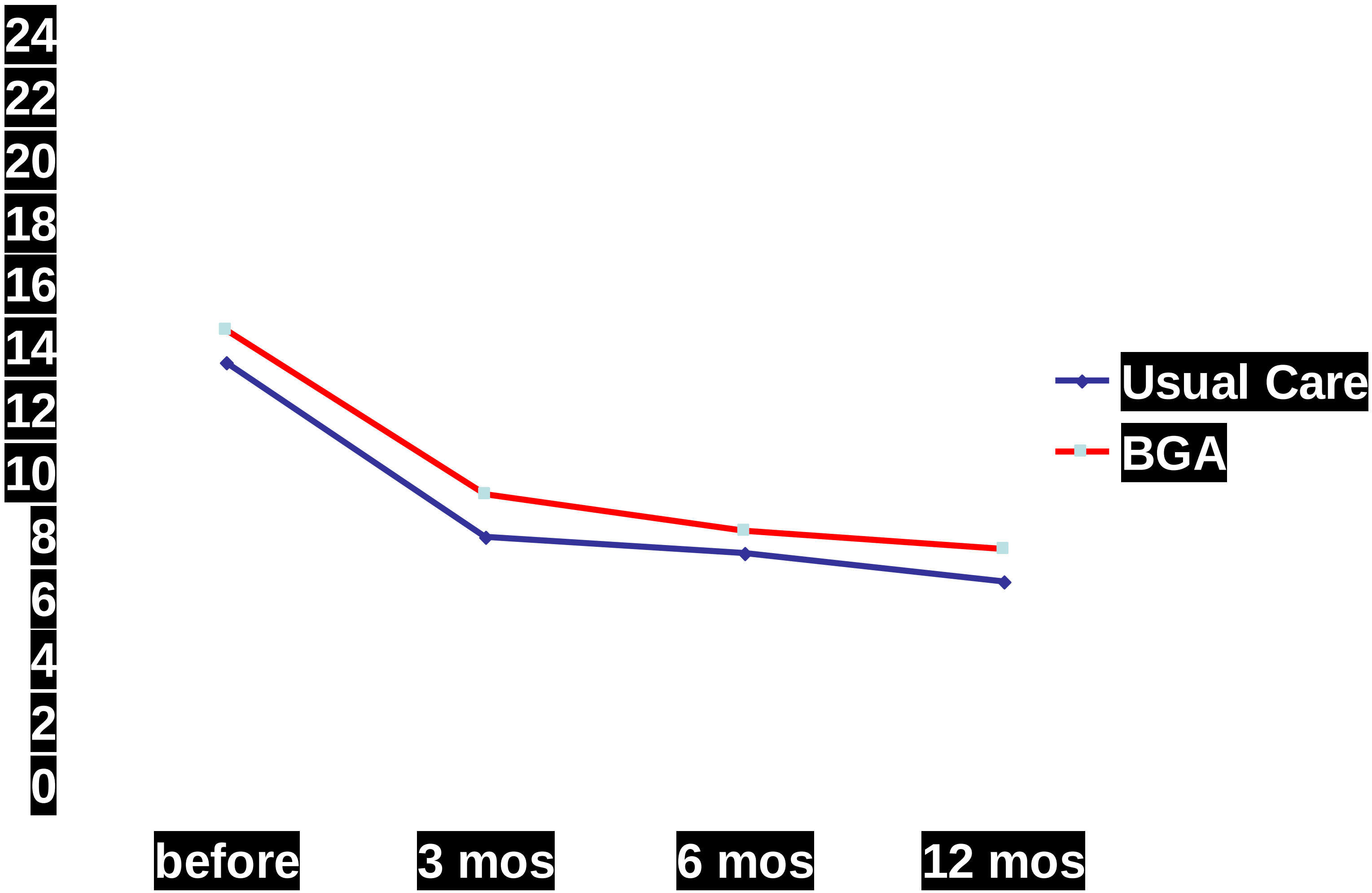
Outcome measures

- Global Perceived Recovery
- Physical Functioning
- Pain
- Fear of movement (re-injury)

Global perceived effect



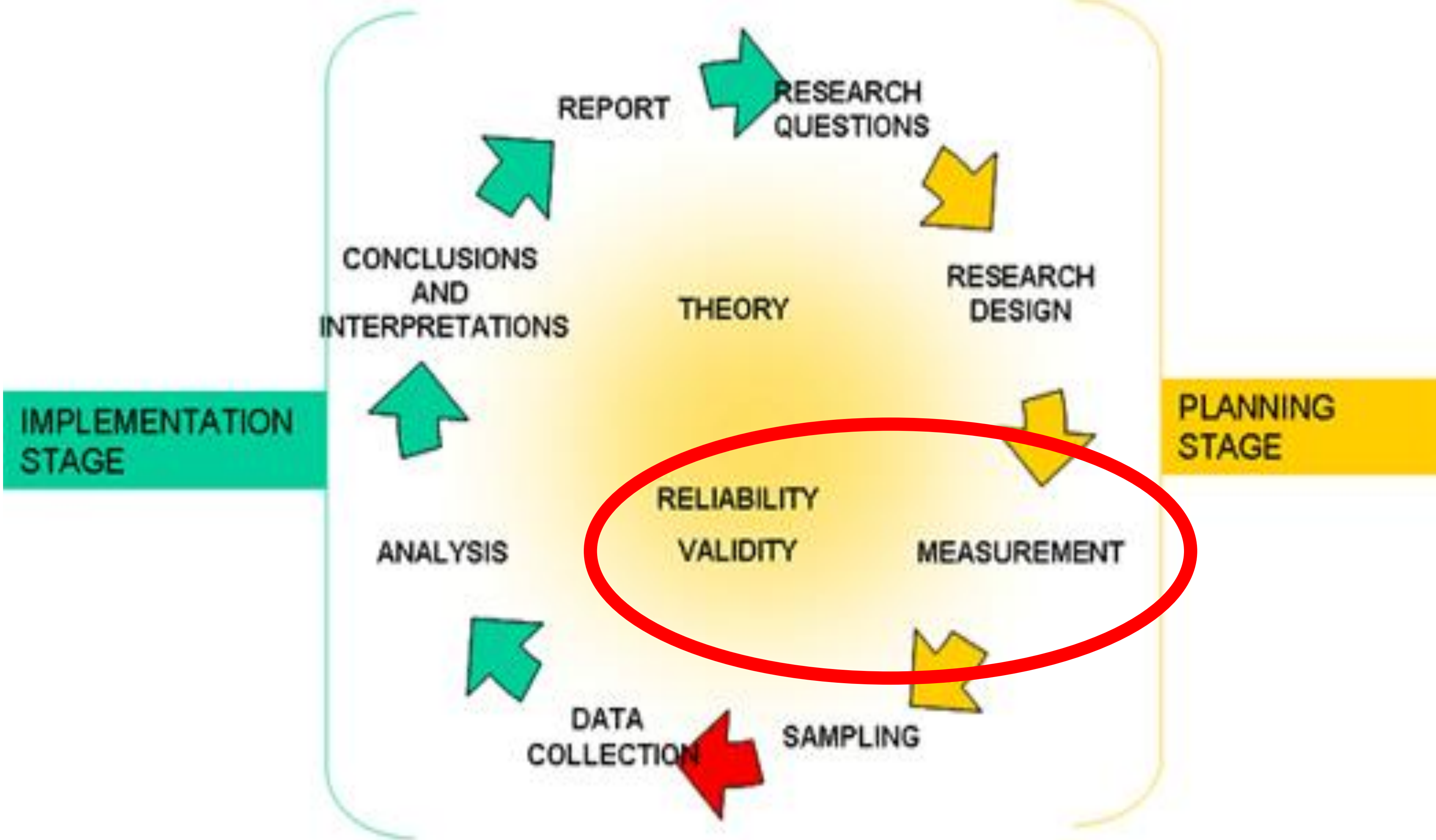
Physical Functioning (RDQ)



6 PROMs to measure physical functioning

Table 3
The variance components and indexes

Questionnaire
RDQ-24
MRDQ
RDQ-18
SF-36 PhF
SF-36 RLPh
MC

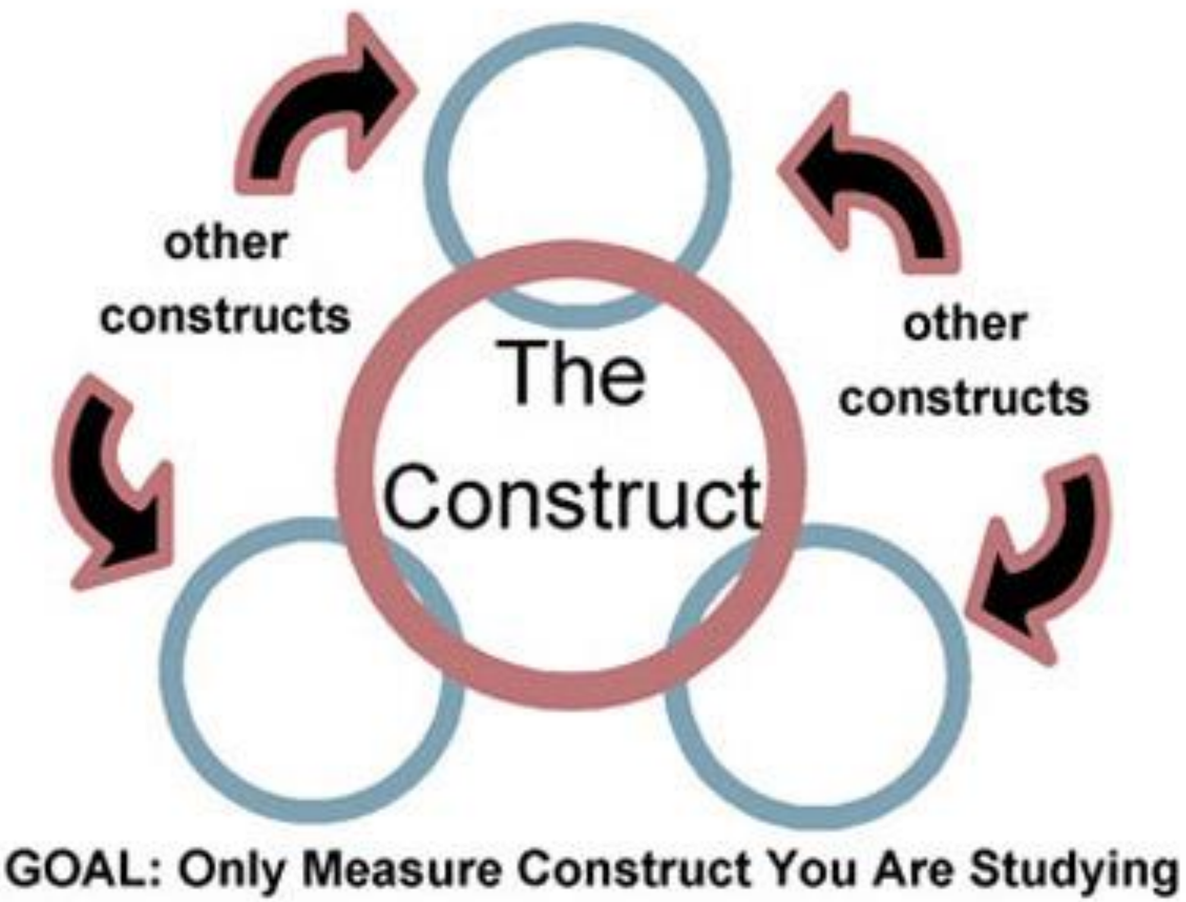

REPORT
RESEARCH QUESTIONS
RESEARCH DESIGN
THEORY
CONCLUSIONS AND INTERPRETATIONS
IMPLEMENTATION STAGE
PLANNING STAGE
RELIABILITY
VALIDITY
MEASUREMENT
ANALYSIS
DATA COLLECTION
SAMPLING


CONFUSED
BY
TOO
MANY
CHOICES


confusion
bemusement
bewilderment
tangle
bafflement
befuddlement
bamboozlement
discombobulation
puzzlement
perplexity
maze
confusing
fog
unclear
muddle
whirl
distraction
head-scratching
confusedness
bewilderedness
mystification
flummoxed
unsure

Validity

Validity


other constructs
other constructs
The Construct
GOAL: Only Measure Construct You Are Studying

The concept of validity

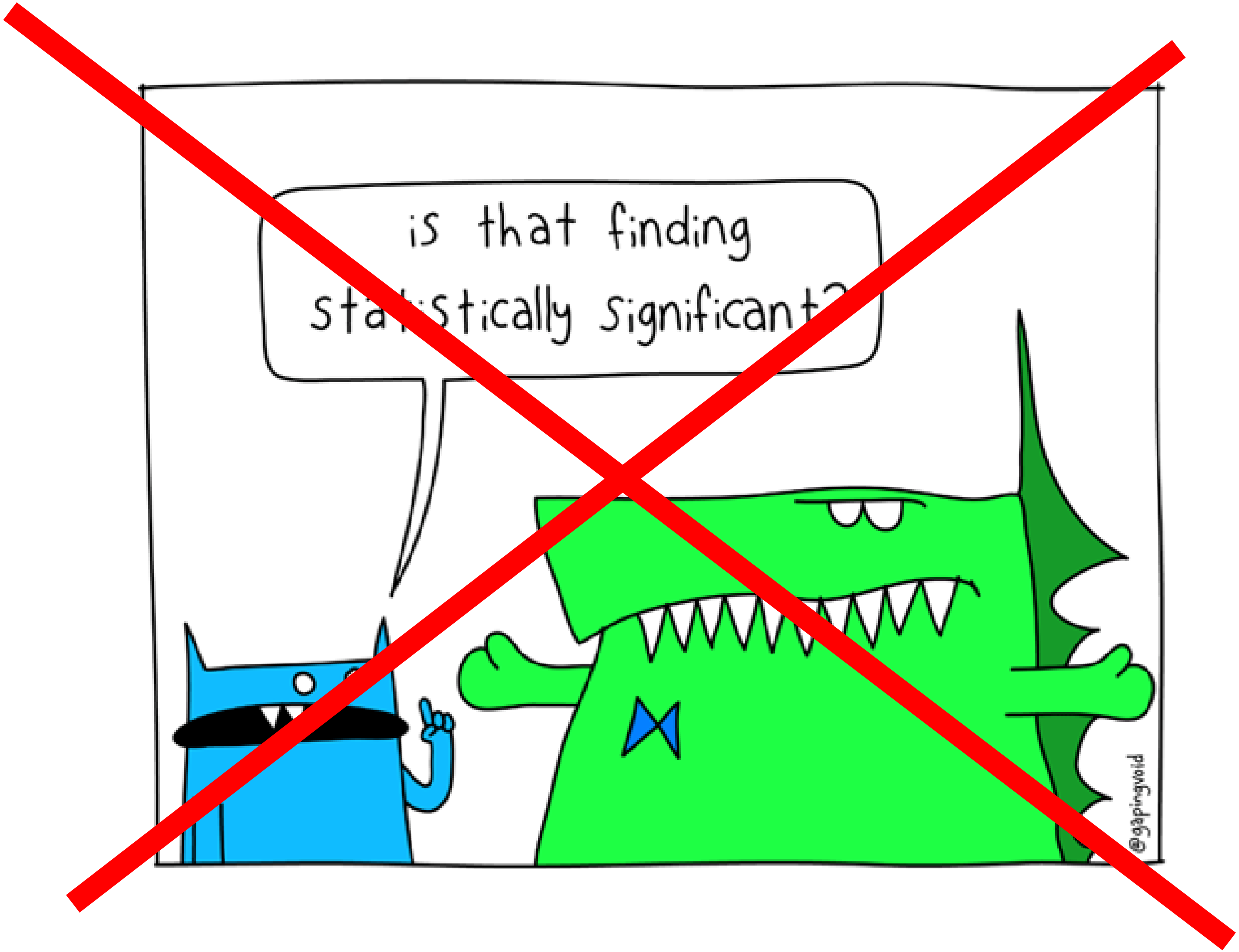
- Knowledge about the construct to be measured
  - Theoretical foundations & conceptual models
- Complexity of the construct
  - Unidimensional vs multidimensional
- Dependency on the situation
  - Target population
- Validation of scores, not measurement instruments
  - Validating the use to which the instrument is put
- Formulation of specific hypotheses
  - Precise theories & models enable strong validation tests
- Validation as a continuous process
  - Often only circumstantial evidence

CONTENT & FACE VALIDITY

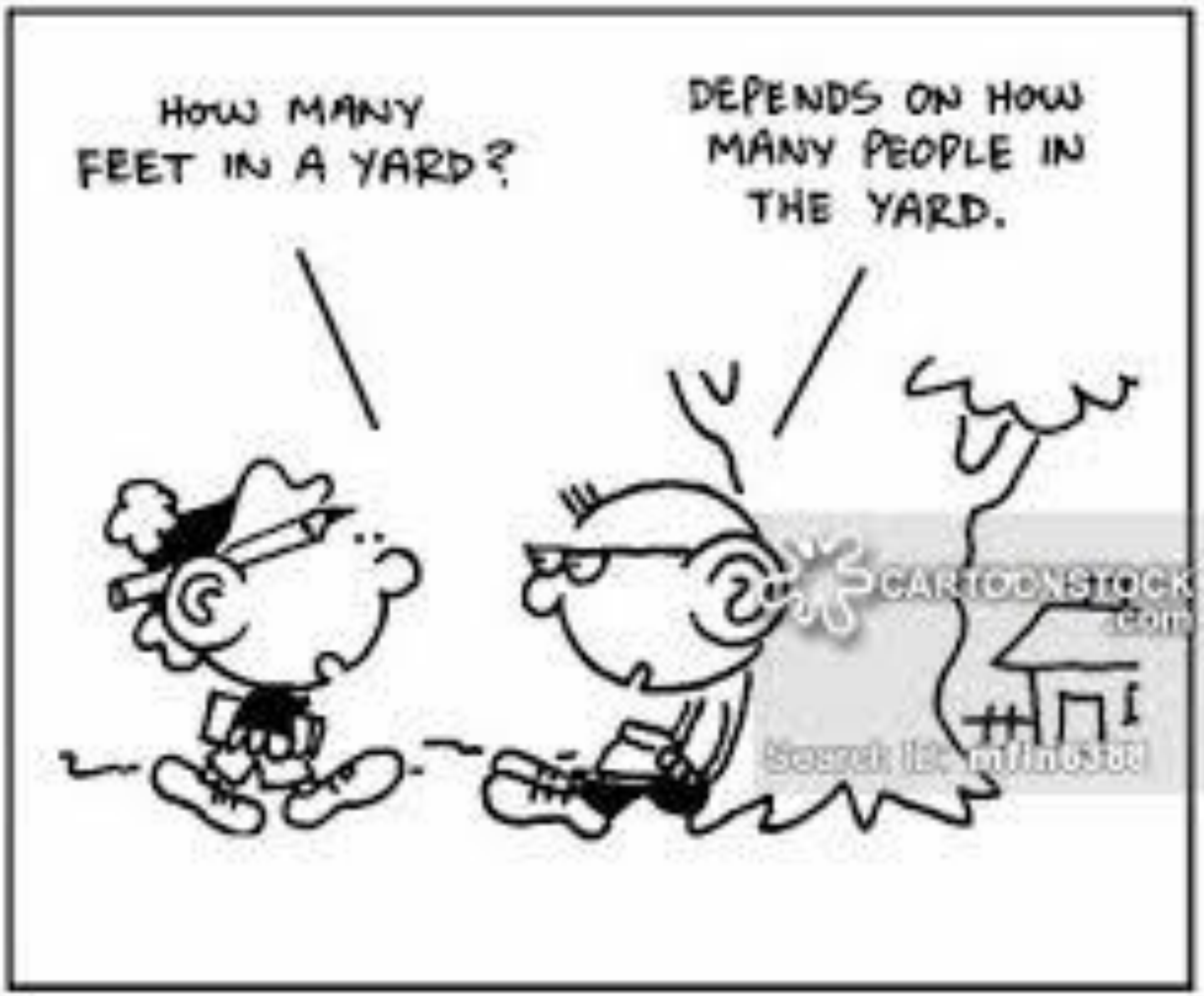
1. Face validity
   - The degree to which an instrument, indeed, looks as though is an adequate reflection of the construct to be measured.

2. Content validity
   - Do all items refer to relevant aspects of the construct?
   - Are all items relevant for study population?
   - Are all items relevant for the purpose of the application of the instrument?

CONTENT & FACE VALIDITY



Validity: do we speak the same language?



Process of content validation: steps to follow

1. Information about construct & situation
   - Specification theoretical models

2. Information about content of instrument
   - Full details, including procedures

3. Select expert panel
   - Independent to prevent ‘over enthusiasm’

4. Assess correspondence between instrument & construct
   - Judgment: sufficiently relevant and comprehensive (also users)

5. Strategy or framework to assess correspondence between instrument & construct

CONTENT OF ITEMS

	RDQ 24	RDQ 18	MC	SF-36 Ph F
Sport (Strenuous)	-	-	?	+
Kneel down / bend	+	+	?	-
Get out of chair	+	+	?	-
Sitting long time	-	-	?	-
Walking	+	+	?	+
Lifting	+	+	?	+

Conclusion

Face validity: all (+)

Content validity: depends…

Framework: an example

Content comparison

ICF category[1]	QL-I	WHO DASII	NHP	SF-36
d450 Walking			1	
d4500 Walking short distances				1
d4501 Walking long distances		1		2
d455 Moving around			2	
d4551 Climbing			2	
d510 Washing oneself	1	1		1
d530 Toileting	1			
d540 Dressing	1	1	1	1
d550 Eating	1	1		
d6309 Preparing meals, unspecified			1	
d640 Doing housework	1	1	1	2
d6509 Caring for household objects		1		

[1]The numbers correspond to various disability (d) categories in the ICF classification

ICF = International Classification of Functioning, QL-I = Quality of Life-Index, WHO DASII = World Health Organisation Disability Assessment Schedule, NHP = Nottingham Health Profile.

Concurrent validity: an example

Cervical Range of Motion (CROM)

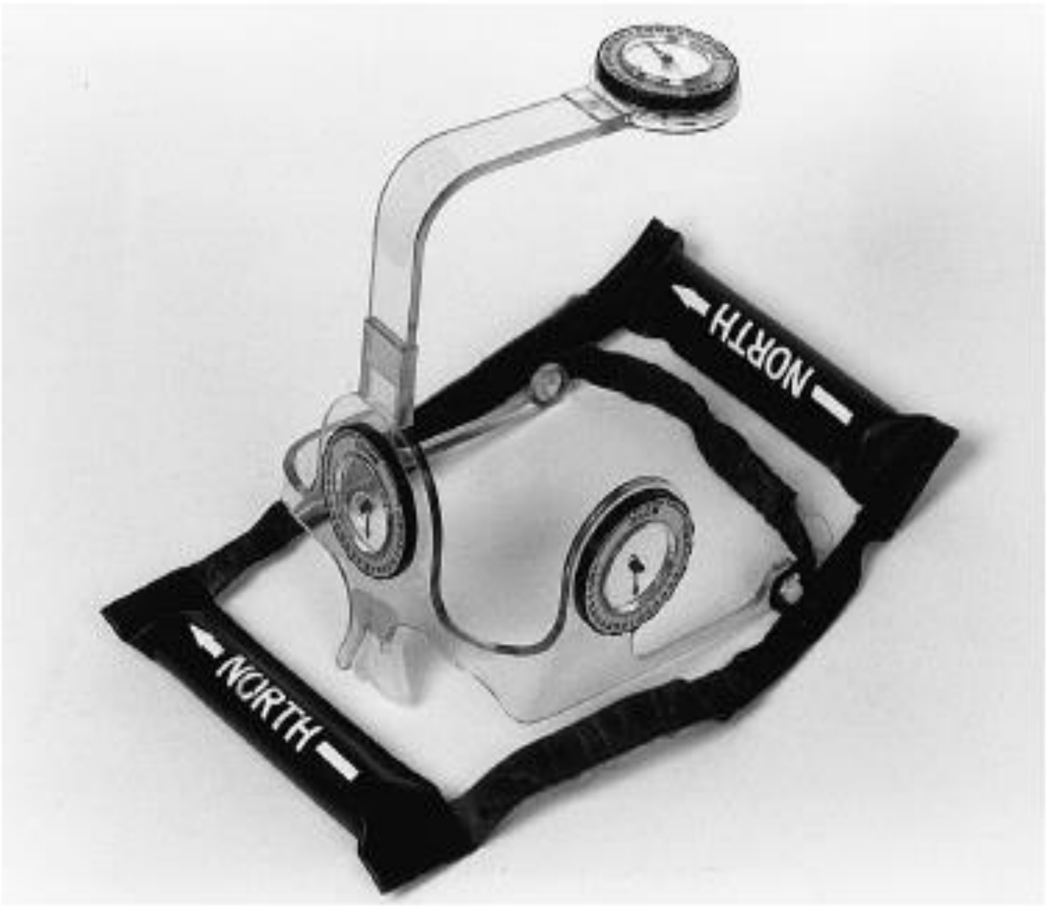
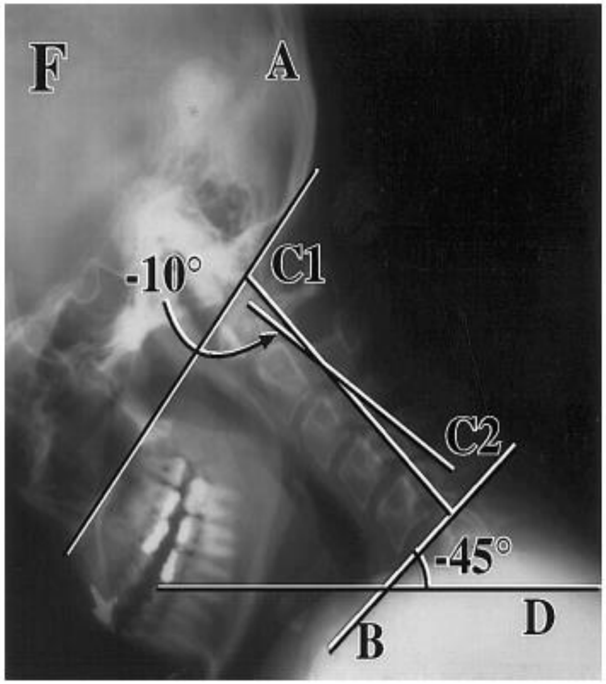
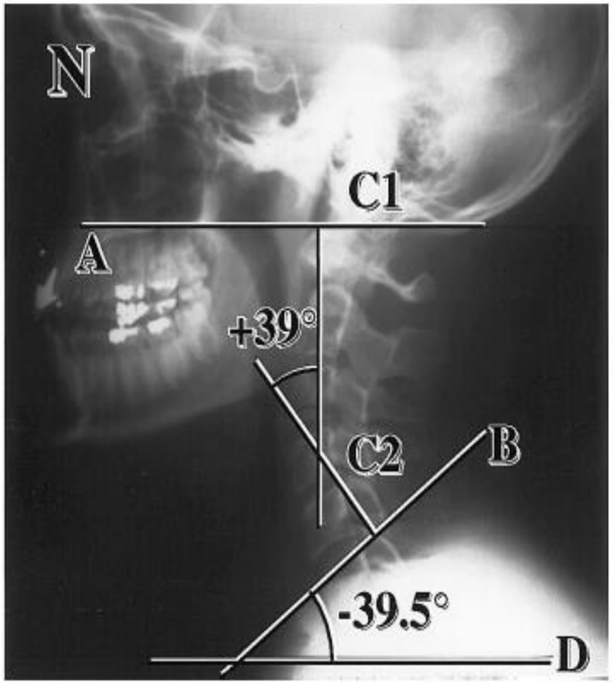


Figure 1. The cervical range of motion (CROM) goniometer.

CRITERION: RADIOGRAPHICS





igure 2. Calculation of cervical ange of motion on radiographs or flexion movement (partici- ant 2).

RESULTS

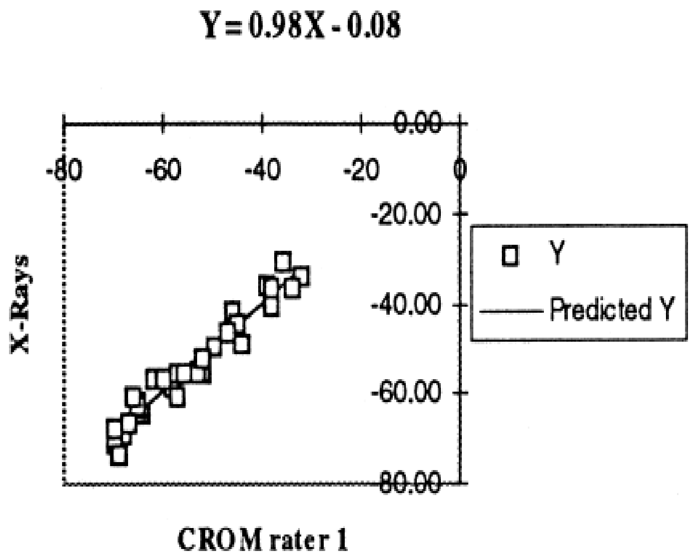


Figure 4. Regression analysis of cervical range of motion on radiographs for flexion.

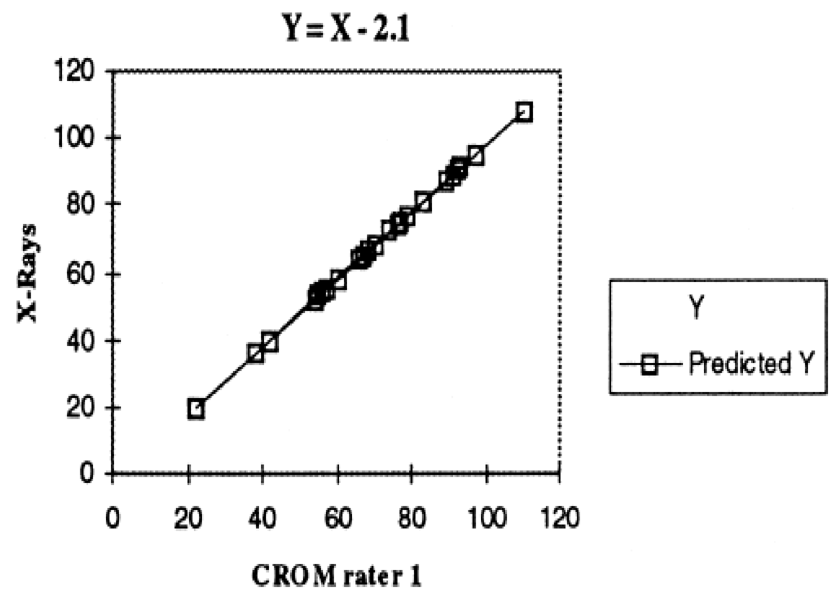


Figure 5. Regression analysis of cervical range of motion on radiographs for extension.

Statistical parameters

Level of measurement		Same units	Statistical parameter
Gold standard	**Measurement instrument**		
dichotomous	dichotomous	yes	sensitivity and specificity
	ordinal	n.a.	ROC
	continuous	n.a	ROC
ordinal	ordinal	yes	weighted kappa
		no	Spearman's r[1] or other measure of association
	continuous	n.a	ROCs[2]/Spearman's r
continuous	continuous	yes	Bland and Altman limits of agreement or ICC[3]
		no	Spearman's r or Pearson's r

[1] r = correlation coefficient; [2] ROCs: for an ordinal gold standard a set of ROCs may be used, dichotomising the instrument by the various cut-off points; [3]ICC – Intraclass Correlation Coefficient

Construct validity: hypotheses testing

- The degree to which scores of an instrument are consistent with hypotheses

hypotheses testing: steps to follow

1. Describe construct to be measured
   - Detailed & conceptual model
2. Formulate hypotheses about expected relationships
   - related constructs or unrelated constructs
   - expected differences between sub-groups of patients
3. Describe measurement instruments of comparator!!
4. Gather empirical data
5. Assess consistency of results and hypotheses
6. Discus observed findings
   - rival theories or alternative explanations

Research Report

Responsiveness to Change of 10 Physical Tests Used for Patients With Back Pain

Liv Inger Strand, Bodil Anderson, Hildegunn Lygren, Jan Sture Skouen, Raymond Ostelo, Liv Heide Magnussen

L.I. Strand, PT, PhD, is Professor, Department of Public Health and Primary Health Care, Physiotherapy Research Group, University of Bergen, Kalfarveien 31, 5018 Ber-

Construct validity

Construct validity (baseline scores)

1. The scores of physical tests of activities were expected to be moderately correlated ($.60 > r \geq .30$) with scores of self-report questionnaires of functioning.
2. The scores of physical tests of body functions were expected to be at least weakly correlated ($.30 > r \geq .20$) with scores of self-report questionnaires of functioning.
3. Scores of all physical tests were expected to be more highly correlated with scores of the Hannover Functional Ability Questionnaire than with scores of the Roland-Morris Disability Questionnaire.
4. Scores of the Back Performance Scale were expected to be most highly correlated with the scores of the self-report questionnaires.

Physical Tests	Baseline Scores	
	FFbH-R[b]	RMDQ[c]
Body functions		
Biering-Sørensen test	−.16	−.33**
Spondylometry	−.37**	−.26**
GPE[d] flexibility subscale	.06	.09
Lateral flexion test	−.40**	−.24*
Fingertip-to-floor test	.20*	<.01
Loaded reach test	−.23*	−.10
Activities		
PILE[e]	−.44**	−.32**
Lift test	−.42**	−.38**
15-m walk test	−.40**	−.37**
Back Performance Scale	.56**	.44**

Validity & Reliability

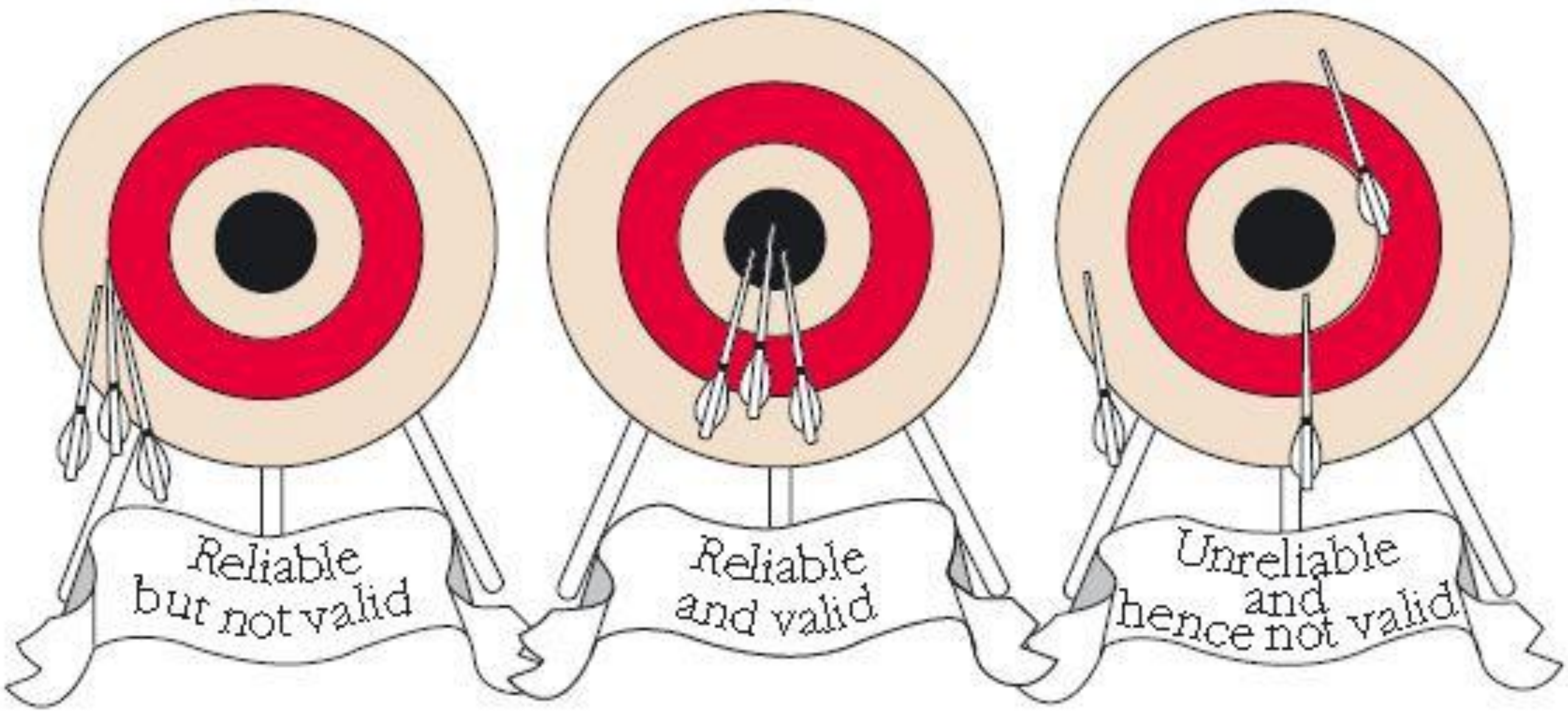


Figure 5.1 *Reliability and validity. (Source: Open University, 1979,* Classification and Measurement, *DE304, Block 5, The Open University, Milton Keynes, p. 68)*

6 PROMs to measure physical functioning

Table 3
The variance components and indexes

Questionnaire	Between subject variance	Within-subject variance		ICC (95% CI)
		Between measures	Residual	
RDQ-24	11.152	0.596	3.257	0.74 (0.51–0.87)
MRDQ	11.520	0.512	2.708	0.78 (0.57–0.89)
RDQ-18	7.868	0.271	2.317	0.75 (0.55–0.87)
SF-36 PhF	185.660	25.561	98.442	0.60 (0.28–0.79)
SF-36 RLPh	121.992	92.469	532.531	0.16 (0–0.45)
MC	83.597	237.373	289.832	0.14 (0–0.40)

Reliability in formula...

$$\text{Reliability} = \frac{\text{var}_{\text{between persons}}}{\text{var}_{\text{between persons}} + \text{error}}$$

$sqrt\ error$ = standard error of measurement (SEM)

Reliability in formula…

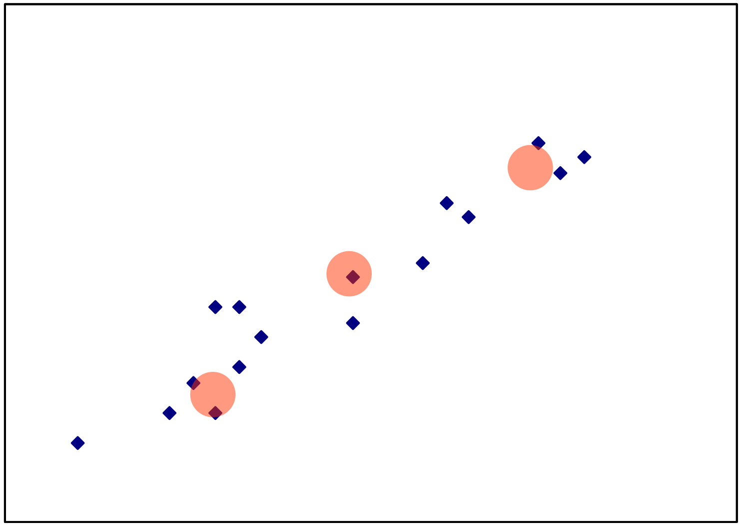
$$\text{Reliability} = \frac{\text{var}_{\text{between persons}}}{\text{var}_{\text{between persons}} + \text{error}}$$

Suppose: error = 1 kg
Adults (range in weight: 50 to 100 kg)
Babies (range in weight: 3 to 5 kg)

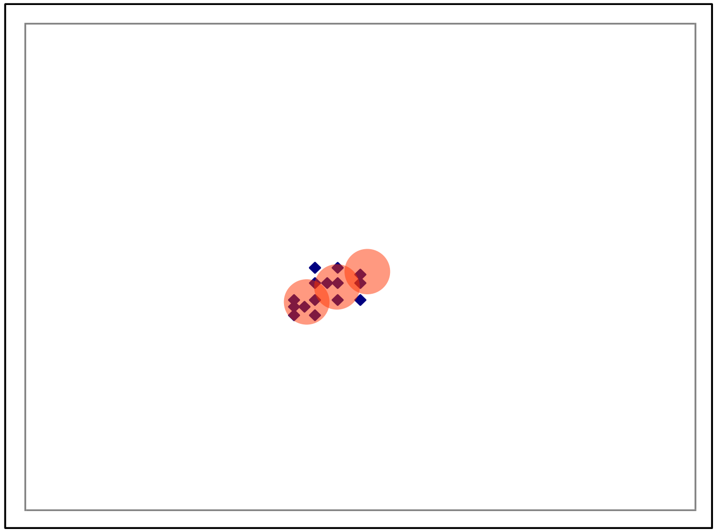
- Reliability Adults = 50 / 50 + 1 = 0,98
- Reliability Babies = 2 / 2 +1 = 0,67

Reliability graphically …

ICC=0.98



ICC=0.67



ICC = Intra Class Correlatiecoëfficiënt

Measurement error and change

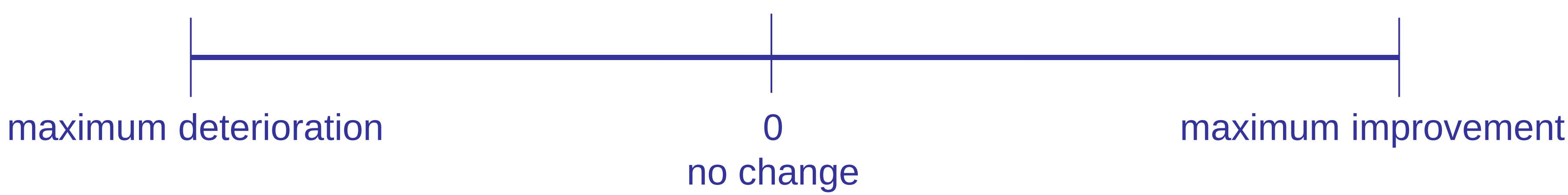
- Linking Smallest Detectable Change (SDC) to Minimal Important Change (MIC)

- Main focus now on interpretation of change scores in individual patients

‘Real’ change

Only change larger than the measurement error can be considered **‘real’** change

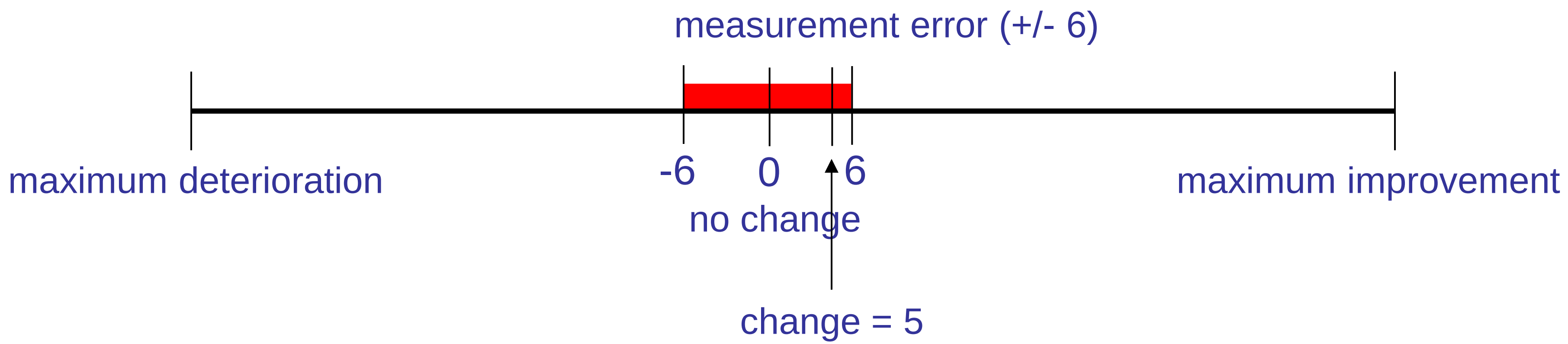
Example



Example 1

Change score = 5 points

Measurement error = 6 points

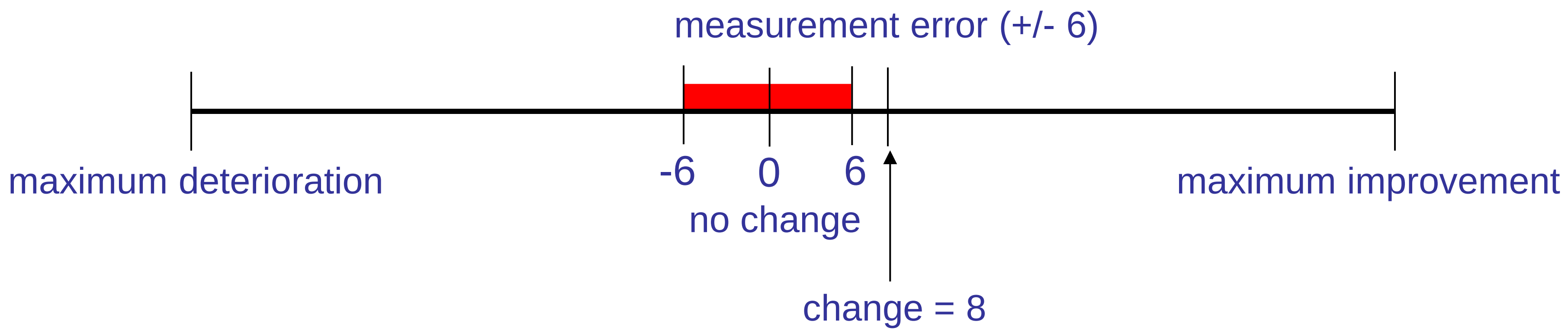


A change of 5 points can NOT be distinguished from no change because of measurement error

Example 2

Change score = 8 points

Measurement error = 6 points



A change of 8 points CAN be considered ‘real’ change

‘Real’ change

Only change larger than the measurement error can be considered ‘real’ change (statistically significant change)

‘real’ change is the smallest change in score that can be detected beyond measurement error

This is called Smallest Detectable Change (SDC)

Smallest Detectable Change

- SDC is a parameter of measurement error
- Should be measured in persons who have NOT changed (stable persons)
- Test-retest design

Smallest Detectable Change (some examples)

Table 3

The variance cc

Questionnaire	SEM (95% CI)	SEM (%)[a] (95% CI)	MDC[b] (95% CI)	MDC (%)[c] (95% CI)
RDQ-24	2.0 (1.5–2.9)	8.2 (6.3–12.1)	5.4 (4.2–8.0)	22.5
MRDQ	1.8 (1.4–2.6)	7.2 (5.6–10.4)	5.0 (3.9–7.2)	21.7
RDQ-18	1.6 (1.2–2.0)	8.9 (6.7–11.1)	4.5 (3.3–5.5)	25.0
SF-36 PhF	11.1 (8.2–17.4)	11.1 (8.2–17.4)	30.9 (22.7–48.2)	30.9
SF-36 RLPh	25.0 (22.8–27.4)	25.0 (22.8–27.4)	69.3 (63.2–75.9)	69.3
MC	23.0 (14.0–61.2)	23.0 (14.0–61.2)	63.6 (38.8–100)	63.6

International Journal of Behavioral Medicine
2007, Vol. 14, No. 4, 242–248

Assessing Pain and Pain-Related Fear in Acute Low Back Pain: What Is the Smallest Detectable Change?

Raymond W. J. G. Ostelo, Ilse J. C. M. Swinkels-Meewisse, Dirk L. Knol, Johan W. S. Vlaeyen, and Henrica C. W. de Vet

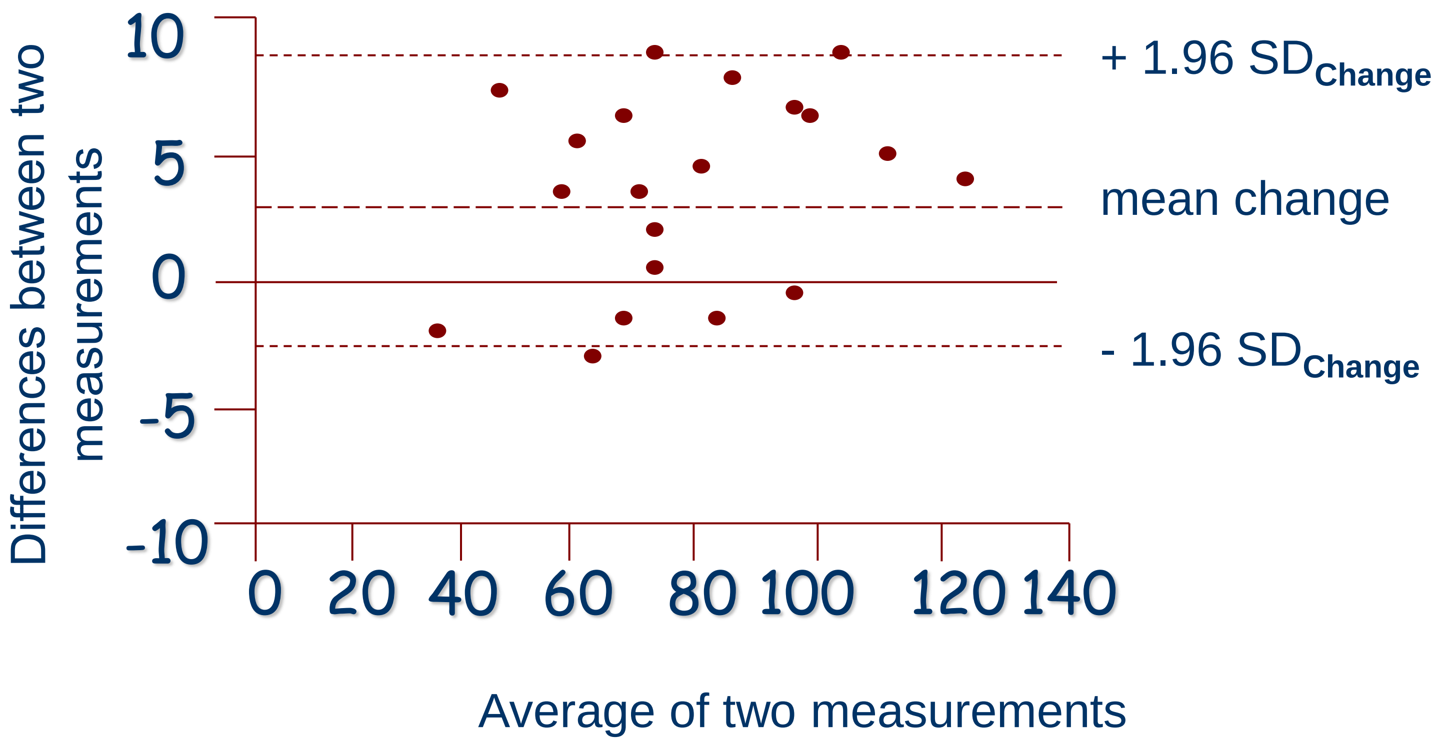
Table 4. *Agreement Parameters (n = 176)*

Questionnaire	Variance Between-Subjects		Variance Within-Subjects		SEM (95% CI)	SEM (%)	SDC	SDC (%)
	Score		*Between Measures*	*Residual*				
Pain (VAS)	0–100	486.156	6.644	163.740	13.1 (11.7, 14.8)	13.1	36.2 (32.4, 41.0)	36.2
TSK total	17–68	38.686	0.000	11.068	3.3 (3.0, 3.7)	6.5	9.2 (8.4, 10.3)	18.0
TSK “harm”	6–24	10.268	0.006	3.108	1.8 (1.6, 2.0)	10.0	4.9 (4.4, 5.5)	27.2
TSK “activity avoidance”	7–28	11.027	0.000	3.849	2.0 (1.8, 2.2)	9.5	5.4 (4.9, 6.1)	25.7
FABQ physical activity	0–24	20.498	0.156	11.461	3.4 (3.1, 3.8)	14.2	9.4 (8.5, 10.6)	39.6
FABQ work	0–42	83.761	0.000	20.864	4.6 (4.1, 5.1)	10.9	12.7 (11.5, 14.1)	30.2

Note. SEM = $\sqrt{\ }$ within-subjects; SEM (%) is SEM expressed in percentages scale related; SDC = 1.96 × $\sqrt{2}$× SEM, SDC (%) is SDC expressed in percentages of scale range.

Smallest Detectable Change

Smallest Detectable Change (SDC) is conceptually equivalent to the limits of agreement (Bland and Altman plot)



Terminology: SDC versus SDD

- Smallest Detectable Change is about changes within persons over time
- Smallest Detectable Difference is about differences between persons (or observers)

Important change

- It is not self-evident that ‘real’ change indicate an *important* change from the patients’, clinicians’ or societal perspective

- A measure of important change = Minimal Important Change (MIC)

- SDC and MIC are different concepts !

METHODS FOR DETERMINING THE MIC

- Data driven methods *Crosby et al. J Clin Epid 2003; 56: 395-407*
  - Distribution-based
    - based on statistical characteristics of the instrument or the population
  - Anchor-based
    - Based on an external criterion that indicates the importance of the change
- Consensus based methods

Linking SDC to MIC

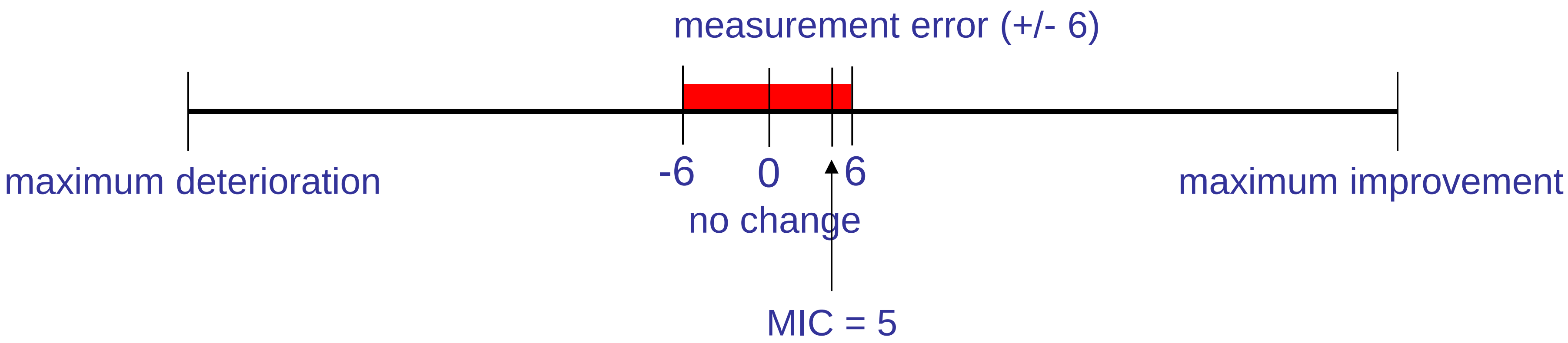
The smallest change that you ***CAN*** detect should be smaller than the smallest change that you ***WANT*** to detect

The SDC should be smaller than the MIC to distinguish important changes from measurement error in individual patients

Example 3

Measurement error (SDC) = 6 points

MIC = 5 points

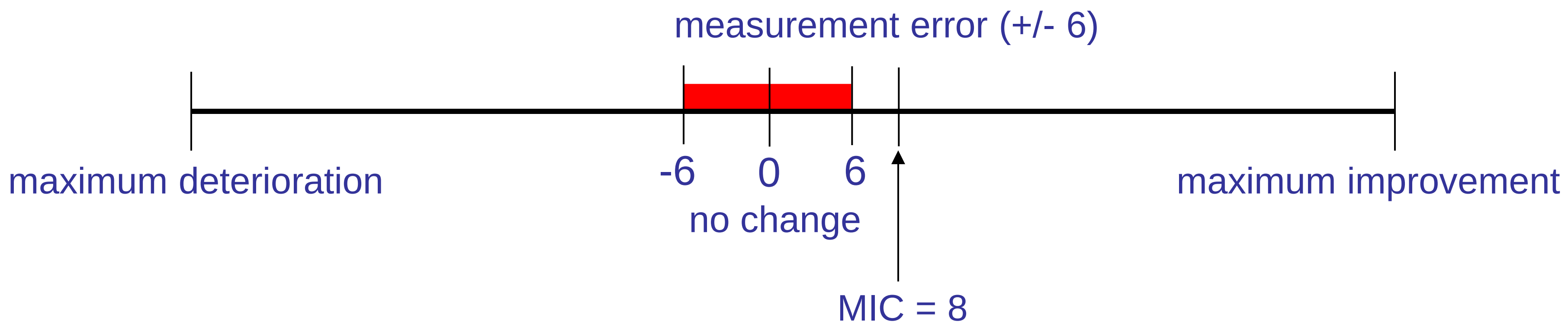


A change as large as the MIC can NOT be distinguished from measurement error

Example 4

Measurement error = 6 points

MIC = 8 points



A change as large as the MIC CAN be distinguished from no change, despite measurement error

Linking SDC to MIC

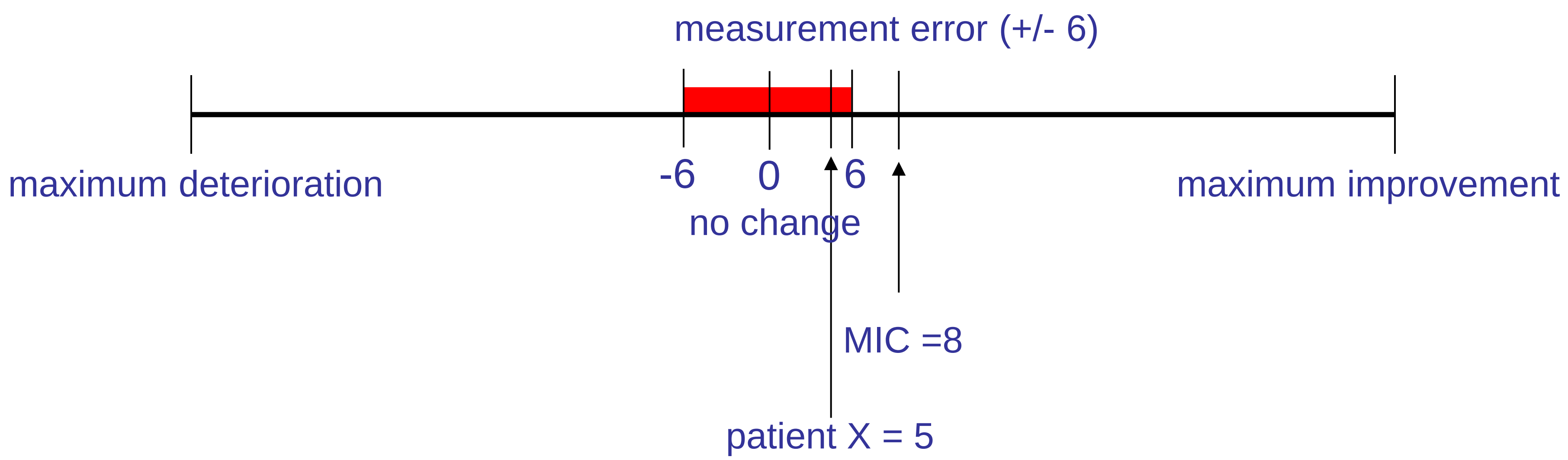
SDC and MIC are two different benchmarks that help to interprete change scores

Example 5

Measurement error = 6 points

MIC = 8 points

Change of patient X = 5 points



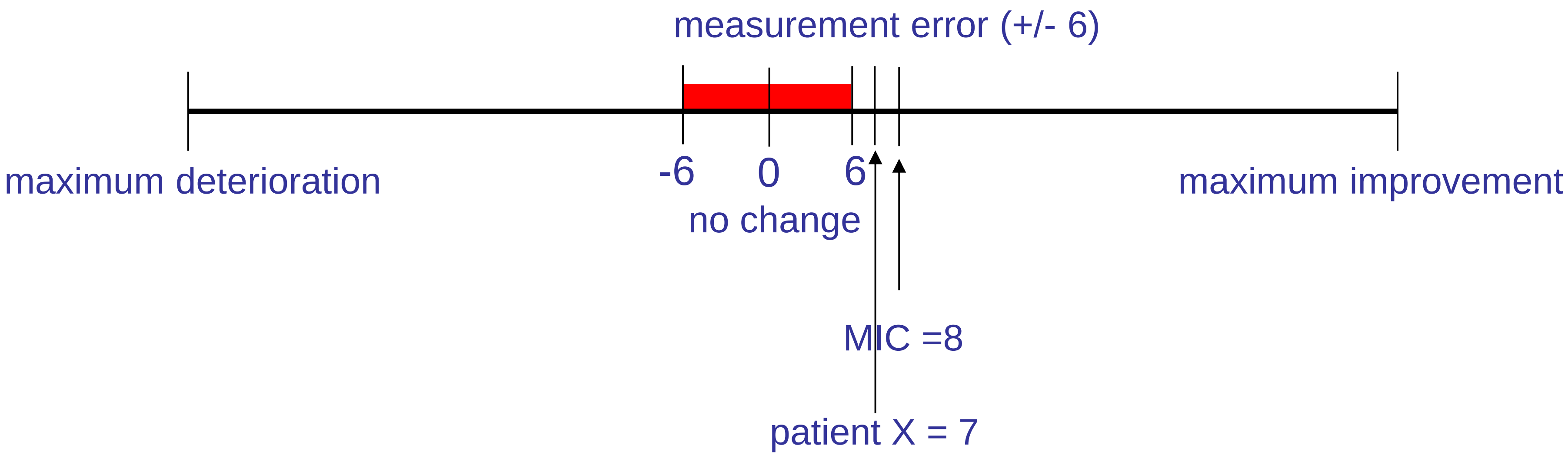
A change of 5 points can NOT be distinguished from no change and is NOT important

Example 6

Measurement error = 6 points

MIC = 8 points

Change of patient X = 7 points



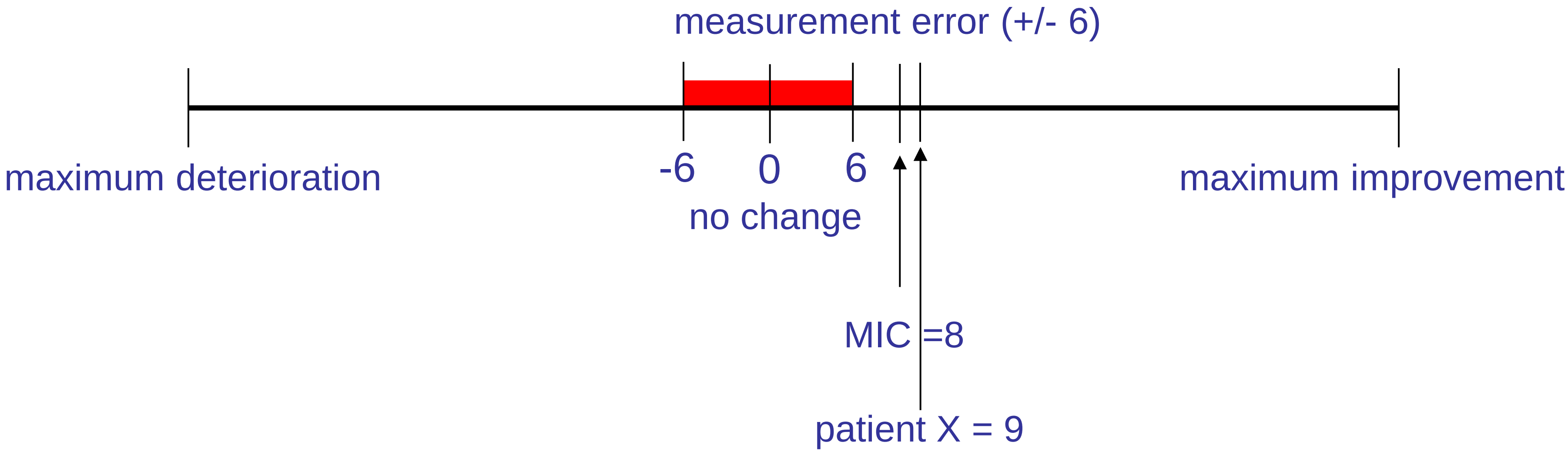
A change of 7 points can be considered 'real' change but NOT important for the patient

Example 7

Measurement error = 6 points

MIC = 8 points

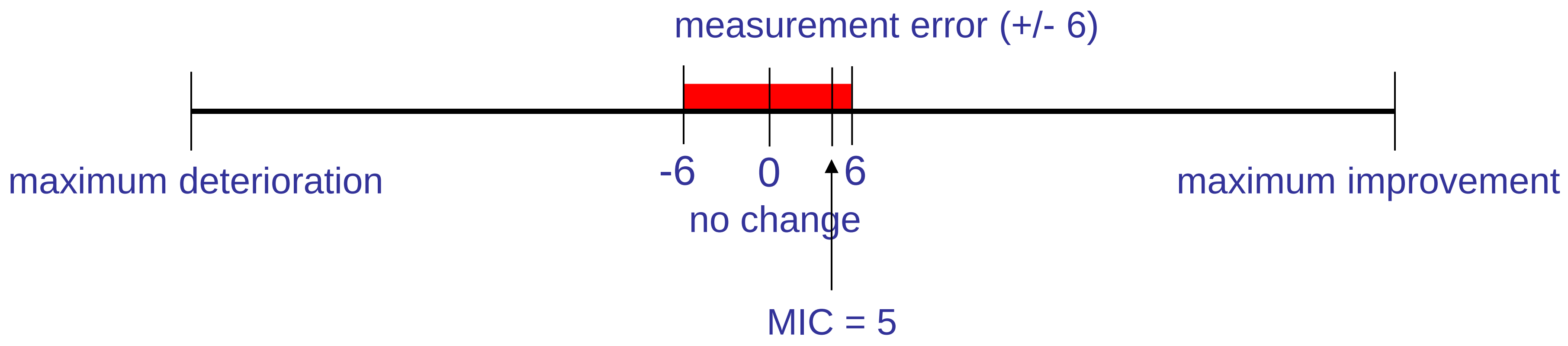
Change of patient X = 9 points



A change of 9 points can be considered 'real' change AND important for the patient

What if SDC > MIC?

If SDC is larger than MIC small but important changes (in an individual person) cannot be distinguished from measurement error



Solution:??

What if SDC > MIC?

- Ma
  wr

- Ma ..

- Ma ….'

Reducing measurement error

1. Increase the number of items in a scale

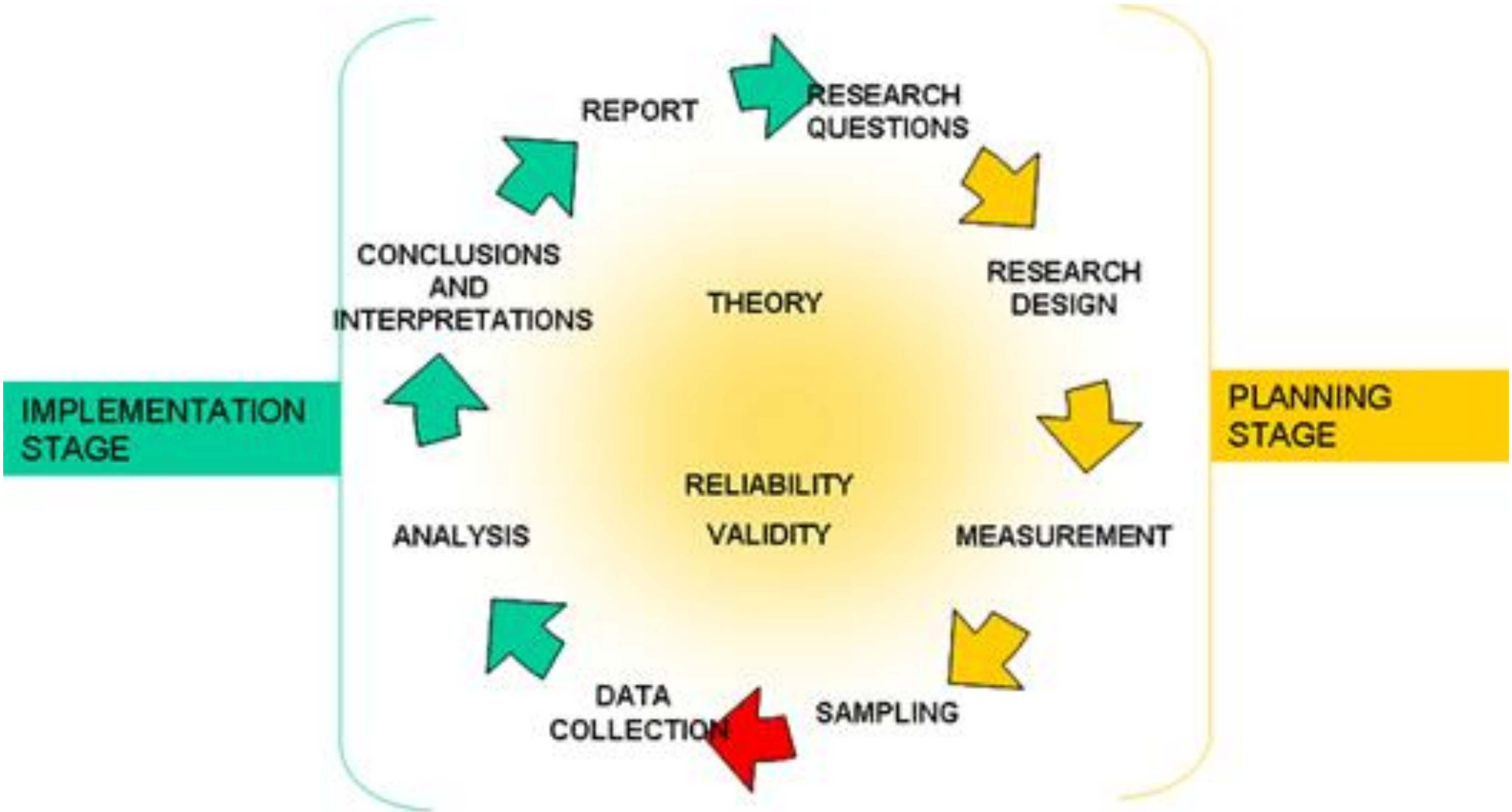
2. Take repeated measurements (k) and average.

   The error variance is divided by k, thus the measurement error is divided by $\sqrt{k}$

Summary

- Validity
  - Face and content validity (no figures or statistical significance but still a structural and valuable approach)
  - Construct validity (hypothesis testing)

- Reliablity
  - Standard error of measurement → SDC
  - Minimal important change
  - SDC and MIC are different concepts!
  - If SDC > MIC, measurement error should be reduced


REPORT
RESEARCH QUESTIONS
RESEARCH DESIGN
MEASUREMENT
SAMPLING
DATA COLLECTION
ANALYSIS
CONCLUSIONS AND INTERPRETATIONS
THEORY
RELIABILITY
VALIDITY
IMPLEMENTATION STAGE
PLANNING STAGE