MEB, Department Seminar, November 9, 2007

Sensitivity to Change of Outcome Measures

Wilmar Igl

Psychologist, PhD

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Introduction (I)

- Measurement of change is frequently applied in clinical research.
- Effects of health changes depend not only on the intervention, but also on the quality of the applied measurement tools.

- Categories of measurement goals (Kirshner & Guyatt, 1985):
  - → Discrimination, e.g. between groups with different morbidity
  - → Prognosis, e.g. prediction of treatment success
  - → Evaluation, e.g. changes caused by interventions

- Regarding evaluation studies sensitivity to change is of primary importance.
- Sensitivity to change/Responsiveness:
  The ability of an instrument to measure change.

Introduction (II)

Weighing machine with a measurement error of ± 3 kg (Norman, 1989):

- Individuals with 60 to 150 kg initial weight
- Intervention with an average weight reduction of 2 kg/week

Sickness Impact Profile (SIP, Bergner et al., 1981)

- Question: "Have you ever attempted suicide?"
- Answer: ☐ yes ☐ no

These measurement tools seem ...

- **adequate** for differentiating between groups (discrimination)
- **inadequate** for measuring change caused by a interventions (evaluation)

I. Theoretical Background

What is a (good) Test?

- Definition (cf. Rost, 2004):
  "A test is a special (psychological) experiment to compare individuals (or groups) and/or items (or aggregated scores)."

- Derivation of test criteria from the general theory of measurement error (cf. classical test theory):

$$X = T + E$$

- Reliability (cf. "precision"):

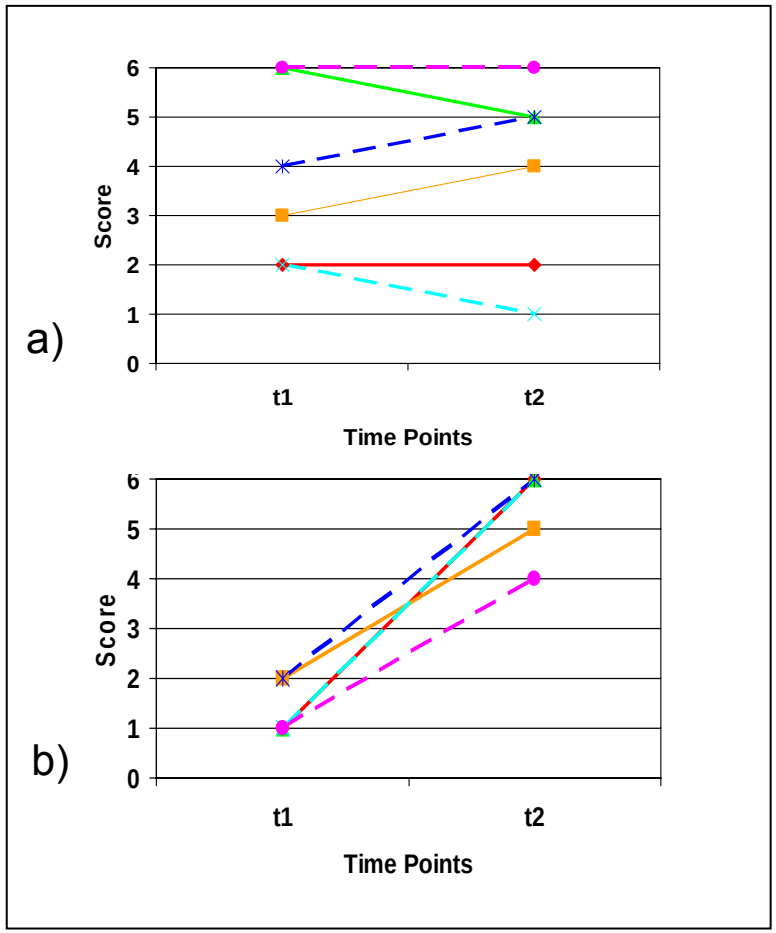
$$Rel = \frac{\sigma^2(T)}{\sigma^2(X)}$$

- Validity (cf. "unbiasedness"):

$$Val = \frac{cov(T, C)}{\sigma(T) \cdot \sigma(C)}$$

Correlational vs. Experimental Approach

- Different methods to measure test criteria are necessary for the measurement of state vs. the measurement of change (cf. Cronbach, 1957).
- **Correlational approach:**
  (cf. measurement of state)
  - → *Inter*-individual differences as ”signal”
  - → *Intra*-individual change as ”noise”
- **Experimental approach:**
  (cf. measurement of change)
  - → *Inter*-individual differences as ”noise”
  - → *Intra*-individual change as ”signal”



Measurement of state: Methods

- **Methods for measuring reliability:**
  - → Test-retest reliability
  - → Parallel-test reliability
  - → Split-half reliability
  - → Internal consistency

- **Methods for measuring validity:**
  - → Criterion validity (e.g. via a valid reference standard)
  - → Construct validity (e.g. via *Multitrait-Multimethod-approach*, Campbell & Fiske, '59)

These methods are usually based on correlational methods!

They are not appropriate for the measurement of (homogenous) change!

Data Example

Measurement	Instrument	Subject	t1	t2	t2-t1	r(tt)	r(tc, t1)	ES=m/sd
State	Test	1	2	2	0	0.89	0.96	0,00
	Test	2	3	4	+1			
	Test	3	6	5	-1			
	Criterion	1	2	1	-1	0.94		0,00
	Criterion	2	4	5	+1			
	Criterion	3	6	6	0			
Measurement	**Instrument**	**Subject**	**t1**	**t2**	**t2-t1**	**r(dd)**	**r(tc, t2-t1)**	**ES=m/sd**
Change	Test	1	1	6	5	NA	0,00	3,75
	Test	2	2	5	3			
	Test	3	1	6	5			
	Criterion	1	1	6	5	NA		4,00
	Criterion	2	2	6	4			
	Criterion	3	1	4	3			

Data Example

Measurement	Instrument	Subject	t1	t2	t2-t1	r(tt)	r(tc, t1)	ES=m/sd
State	Test	1	2	2	0	0.89	0.96	0,00
	Test	2	3	4	+1			
	Test	3	6	5	-1			
	Criterion	1	2	1	-1	0.94		0,00
	Criterion	2	4	5	+1			
	Criterion	3	6	6	0			
Measurement	**Instrument**	**Subject**	**t1**	**t2**	**t2-t1**	**r(dd)**	**r(tc, t2-t1)**	**ES=m/sd**
Change	Test	1	1	6	5	NA	0,00	3,75
	Test	2	2	5	3			
	Test	3	1	6	5			
	Criterion	1	1	6	5	NA		4,00
	Criterion	2	2	6	4			
	Criterion	3	1	4	3			

Data Example

Measurement	Instrument	Subject	t1	t2	t2-t1	r(tt)	r(tc, t1)	ES=m/sd
State	Test	1	2	2	0	0.89	0.96	0,00
	Test	2	3	4	+1			
	Test	3	6	5	-1			
	Criterion	1	2	1	-1	0.94		0,00
	Criterion	2	4	5	+1			
	Criterion	3	6	6	0			
Measurement	**Instrument**	**Subject**	**t1**	**t2**	**t2-t1**	**r(dd)**	**r(tc, t2-t1)**	**ES=m/sd**
Change	Test	1	1	6	5	NA	0,00	3,75
	Test	2	2	5	3			
	Test	3	1	6	5			
	Criterion	1	1	6	5	NA		4,00
	Criterion	2	2	6	4			
	Criterion	3	1	4	3			

Data Example

Measurement	Instrument	Subject	t1	t2	t2-t1	r(tt)	r(tc, t1)	ES=m/sd
State	Test	1	2	2	0	0.89	0.96	0,00
	Test	2	3	4	+1			
	Test	3	6	5	-1			
	Criterion	1	2	1	-1	0.94		0,00
	Criterion	2	4	5	+1			
	Criterion	3	6	6	0			
Measurement	**Instrument**	**Subject**	**t1**	**t2**	**t2-t1**	**r(dd)**	**r(tc, t2-t1)**	**ES=m/sd**
Change	Test	1	1	6	5	NA	0,00	3,75
	Test	2	2	5	3			
	Test	3	1	6	5			
	Criterion	1	1	6	5	NA		4,00
	Criterion	2	2	6	4			
	Criterion	3	1	4	3			

Data Example

Measurement	Instrument	Subject	t1	t2	t2-t1	r(tt)	r(tc, t1)	ES=m/sd
State	Test	1	2	2	0	0.89	0.96	0,00
	Test	2	3	4	+1			
	Test	3	6	5	-1			
	Criterion	1	2	1	-1	0.94		0,00
	Criterion	2	4	5	+1			
	Criterion	3	6	6	0			
Measurement	**Instrument**	**Subject**	**t1**	**t2**	**t2-t1**	**r(dd)**	**r(tc, t2-t1)**	**ES=m/sd**
Change	Test	1	1	6	5	NA	0,00	3,75
	Test	2	2	5	3			
	Test	3	1	6	5			
	Criterion	1	1	6	5	NA		4,00
	Criterion	2	2	6	4			
	Criterion	3	1	4	3			

Measurement of Change: Methods

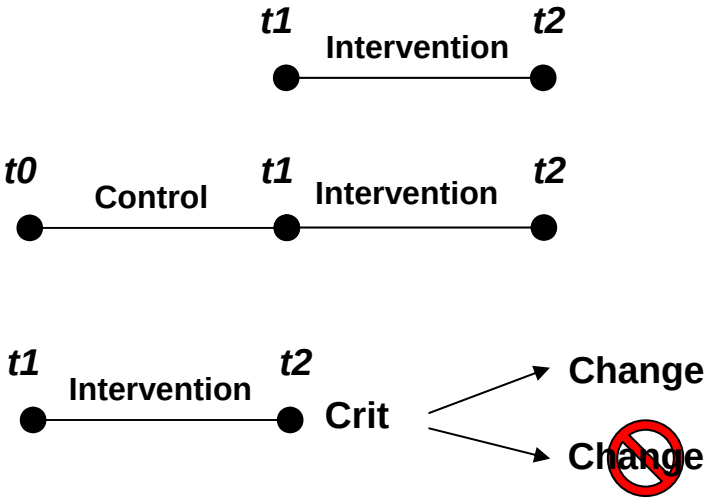
- Definition:
  „Sensitivity to change is defined as the ability of an instrument to measure "true" change of a latent construct.“
- Synonyms: responsiveness, longitudinale validity, evaluative validity

- *At present:*
  Controversial discussion of different definitions and methods (25 definitions, 31 coefficients/designs, Terwee et al., 2003)

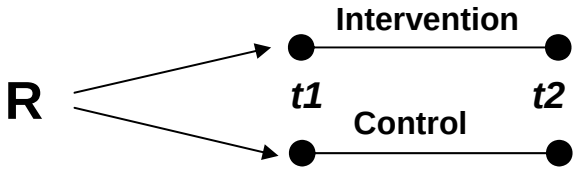
- *Theoretical categories (Terwee et al., 2003):*
  a) *Ability to measure change in general*
  b) *Ability to measure clinical significant change*
  c) *Ability to measure change of a construct of interest*

Study Designs

- Single-group designs
  - → simple pre-post design
  - → pre-post design with baseline
  - → simple pre-post design with external criteria



- Two-group designs
  (with intervention and control group)

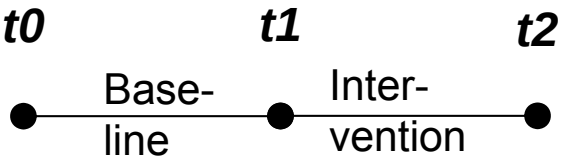


Coefficients

- **Criterion-based coefficients**
  - → Pearson correlation r
  - → Linear regression
  - → AUC (ROC analysis)

- **Distribution-based coefficients**
  - → Statistical tests (*t*- , *F*- , *p-values*)
  - → Norman's coefficients S_{ANOVA} and S_{ANCOVA}
  - → Standardized effect sizes:
    - easy to calculate
    - not dependent on sample size
    - widespread application
    - ... but interpretation depends on specific design and formula

Standardized Effect Sizes



$$SES = \frac{m(t_2) - m(t_1)}{s(t_1)}$$

- SD of values x(t1) at the **beginning of intervention**
- high values if variability is small

$$SRM = \frac{m(t_2) - m(t_1)}{s(t_2 - t_1)}$$

- SD of the difference x(t2) - x (t1) of the **intervention interval**
- high values if variability of differences x(t2) - x(t1) is small

$$GRI = \frac{m(t_2) - m(t_1)}{s(t_1 - t_0)}$$

- SD of differences x(t1) - x(t0) of the **baseline interval**
- high values if variability of differences x(t1) - x(t0) is small
- measuring/confounding change and stability

Using Controlled Trials to Compare Measurement Tools

- Expenditures for clinical trials could be minimized by using better measurement tools.
- Comparative studies to select the best instruments are necessary.
- Designs applying multiple instruments on the same sample are most useful.
- Instruments can be evaluated using descriptive coefficients.
- Sampling error has to be taken into account by calculating confidence intervals or using hypothesis tests.

II. A prospective, comparative study

Igl, W., Zwingmann, C. & Faller, H. (2006). Änderungssensivitität von Fragebogen zur Erfassung der subjektiven Gesundheit – Ergebnisse einer prospektiven vergleichenden Studie [Sensitivity to change of questionnaires measuring subjective health - results of a prospective comparative study]. Rehabilitation 45(4), 232-242.

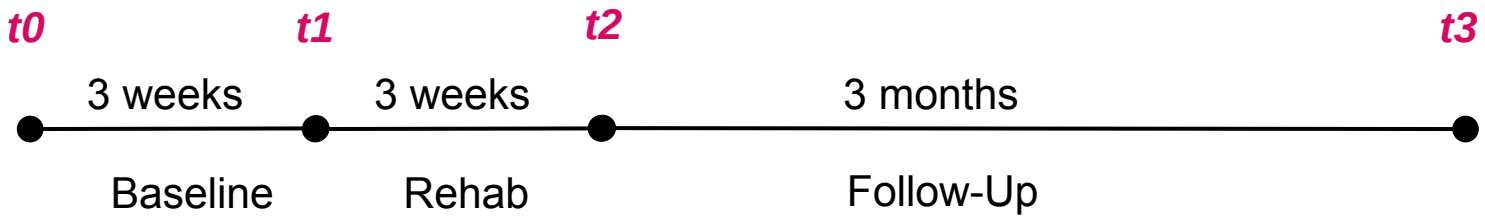
Sample

	Total	Orthopedic/ rheumatic Rehabilitation	Cardiac Rehabilitation
N(Participants)	1145 (100%)	745 (65%)	400 (35%)
Sex	N(F) = 42% N(M) = 58%	N(F) = 53% N(M) = 47%	N(F) = 22% N(M) = 78%
Age	M = 50 years SD = 8.6 years	M = 49 years SD = 8.9 ýears	M = 51 years SD = 7.9 years

Sample and Study Design (I)

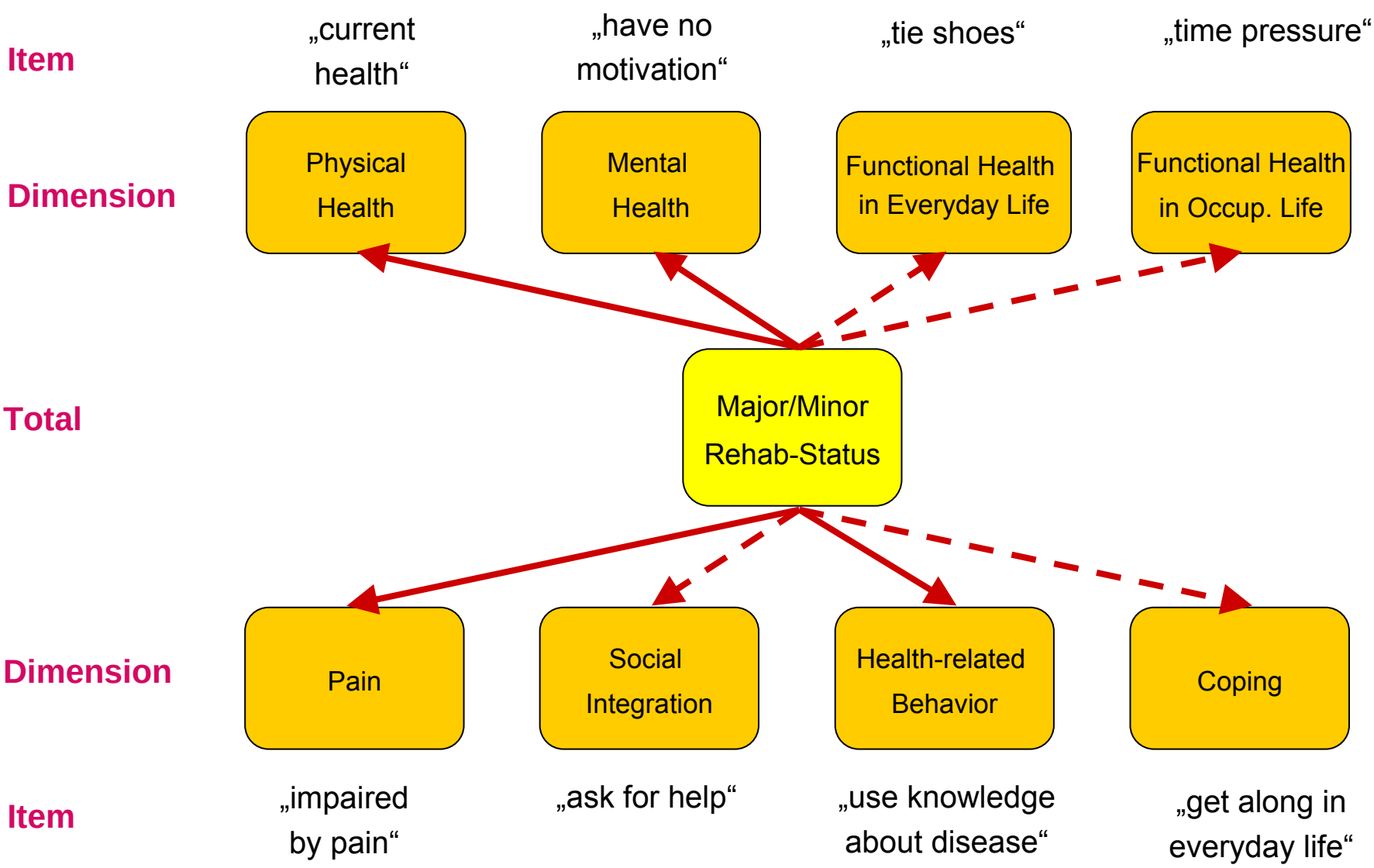
- Sample:
  - → N = 745 patients, orthopedics/rheumatology (ICD, M00-M99),
  - → N = 400 patients, cardiology (ICD, I00-I99)
- Treatment: „Medical Rehabilitation“ (usual care)

- Prospective, comparative one-group design:



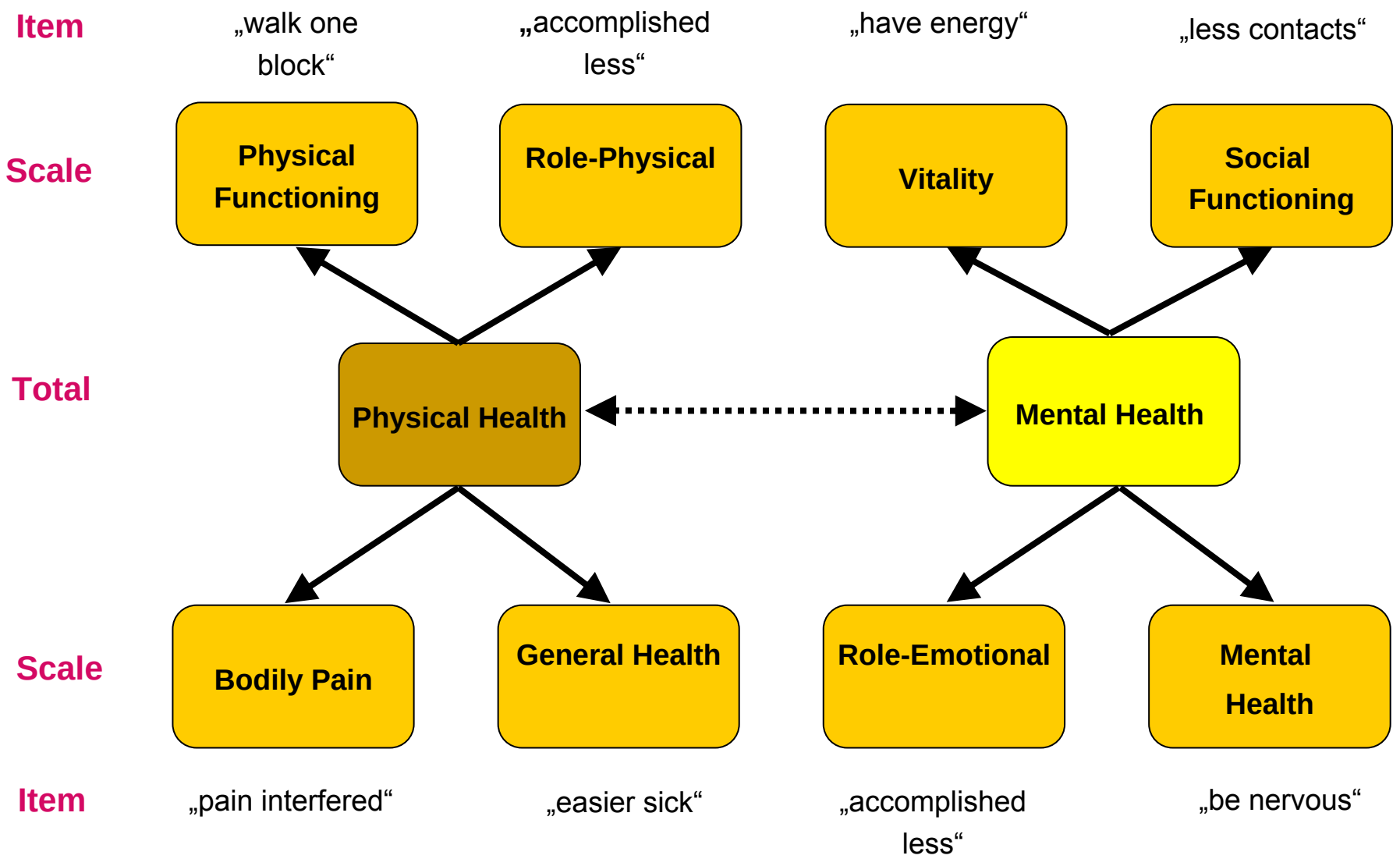
t0	t1	t2	t3
IRES-3.1 SF-36 SCL-90-R (Sc. 1,4,5)	IRES-3.1 SF-36 SCL-90-R (Sc. 1,4,5) SocDat	IRES-3.2 SF-36m SCL-90-R (Sc. 1,4,5) MD Rating	IRES-3.3 SF-36 SCL-90-R (Sc. 1,4,5)

IRES-3* – Measurement Model (scales omitted)



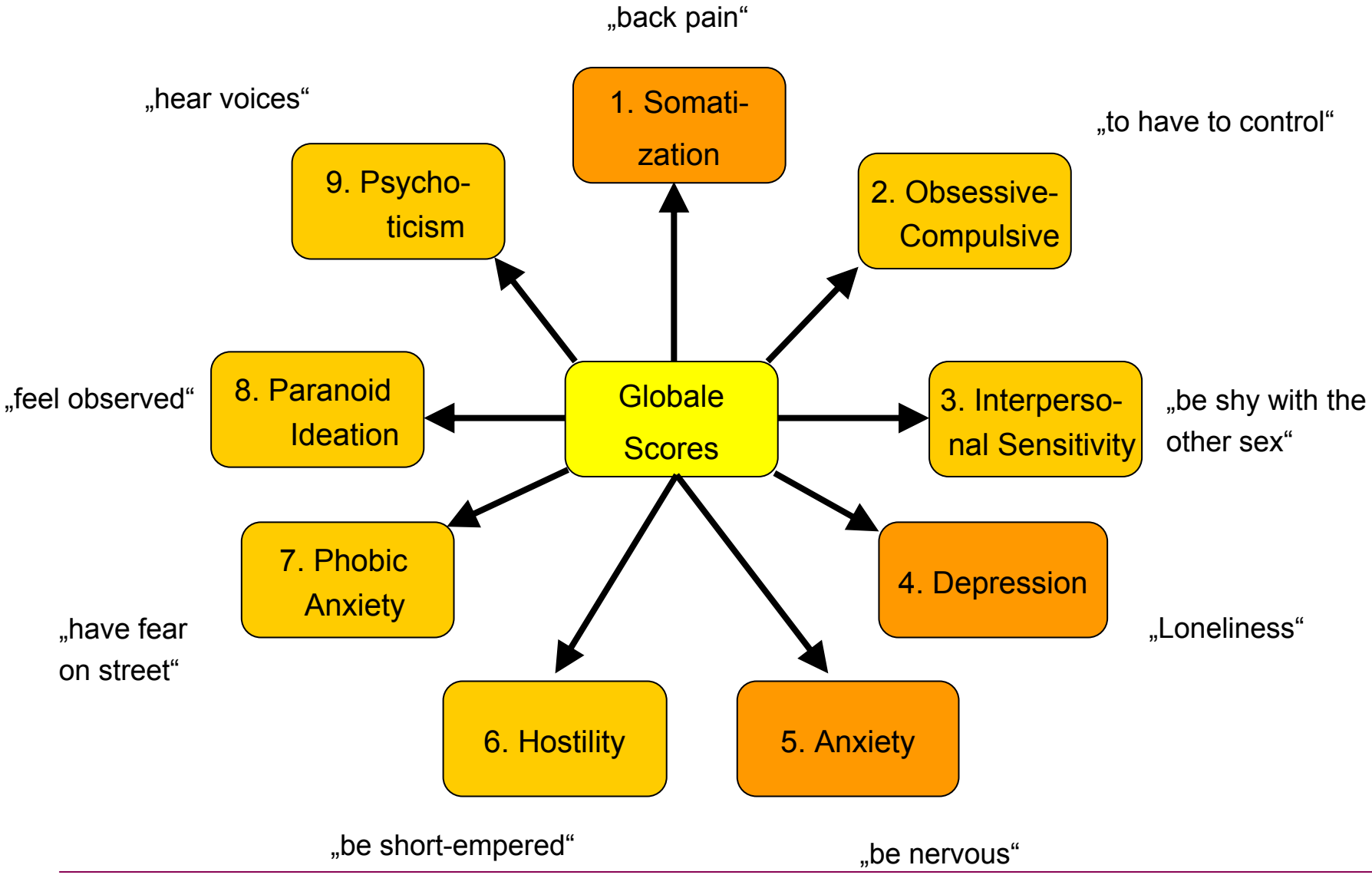
* Bührlen, B., Gerdes, N. & Jäckel, W. H. (2005)

SF-36* – Measurement Model



* Bullinger, M. & Kirchberger, I. (1998)

SCL-90-R* – Measurement Model



* Franke, G. H. (2002)

Study Design (II)

- Independent Variables:
  - → 1st factor: Interventions (baseline, usual care, follow-up)
  - → 2nd factor: Measurement instruments (IRES-3, SF-36, SCL-90-R (1, 4,5), Health-related quality of life (HRQOL) questionnaires)

- Dependent Variables:
  - → 1st end point: Coefficients of sensitivity to change (SRM)
  - → 2nd end point: Coefficients of stability (r_{tt} , Cronbach's α)

- Confounding Variables:
  - → Sequence of questionnaires: complete permutation (ABC, ACB, …, CBA)
  - → Sequence of interventions: not controlled (only empirically!)
  - → Different constructs: theoretical (cross-person model of health-related quality of life) and empirical (principal component analysis) matching of scores

Statistical analysis

- Separate analysis according to diagnosis, measurement intervals, constructs and aggregation levels of scores

- Descriptive evaluation of coefficients based on the distribution of coefficients per diagnostic group and measurement interval

- Sensitivity to change beyond error:
  Estimation of confidence intervals (95%, 2-sided) of SRM coefficients based on non-central *t*-distributions (Smithson, 2002).

- Relative sensitivity to change:
  Estimation of confidence intervals (99%, 2-sided) of *differences of SRM* coefficients based on resampling methods (Hesterberg et al., 2005).

Principal Components Analysis

- **Principal Components Analysis** (PCA) including scores of all questionnaires (Varimax rotation, available case approach, Eigen values > 1)
  - → Similar factor solutions in both diagnostic groups
  - → Score correlate with the expected factors in most cases

- **Principal Component Analysis I** (IRES-3 dimensions, no IRES-3 scales):
  - → 19 scores, N = [299;626]
  - → 2 factors: ”Psyche”, ”Soma” (Var(explained, per factor) ≈ 30% per Factor)
  - → Variance(explained, total) ≈ 62%

- **Principal Component Analysis II** (IRES-3 scales, no IRES-3 dimensions):
  - → 38 scales, N = [297;626]
  - → 6 factors: ”Psyche”, ”Pain”, ”EverydayLife”, ”Social”, ”Behavior”, ”Job” (Var(explained, per factor) < 19%)
  - → Variance(explained, total) ≈ 65%

Analysis of Coefficients

Comparison of effect sizes (SES, SRM, GRI):

- Substantial differences in means: M(SES) < M(SRM) << M(GRI)
- Partially differences in standard deviation: SD(SES) ≈ SD(SRM) < SD (GRI)
- high correlations: r > 0.90

Table: Descriptive Evaluation of SRM coefficients in categories

		SRM		
	Diagnosis	**”small”**	**”medium”**	**”large”**
Admission to Discharge	OR/K	[0.00;0.35]	[0.35;0.70]	[0.70;1.05]
Admission to Follow-Up	OR K	[0.00;0.16] [0.00;0.20]	[0.16;0.33] [0.20;0.40]	[0.33;0.49] [0.40;0.60]
Cohen’s *d*		0.2	0.5	0.8

General Health (I)

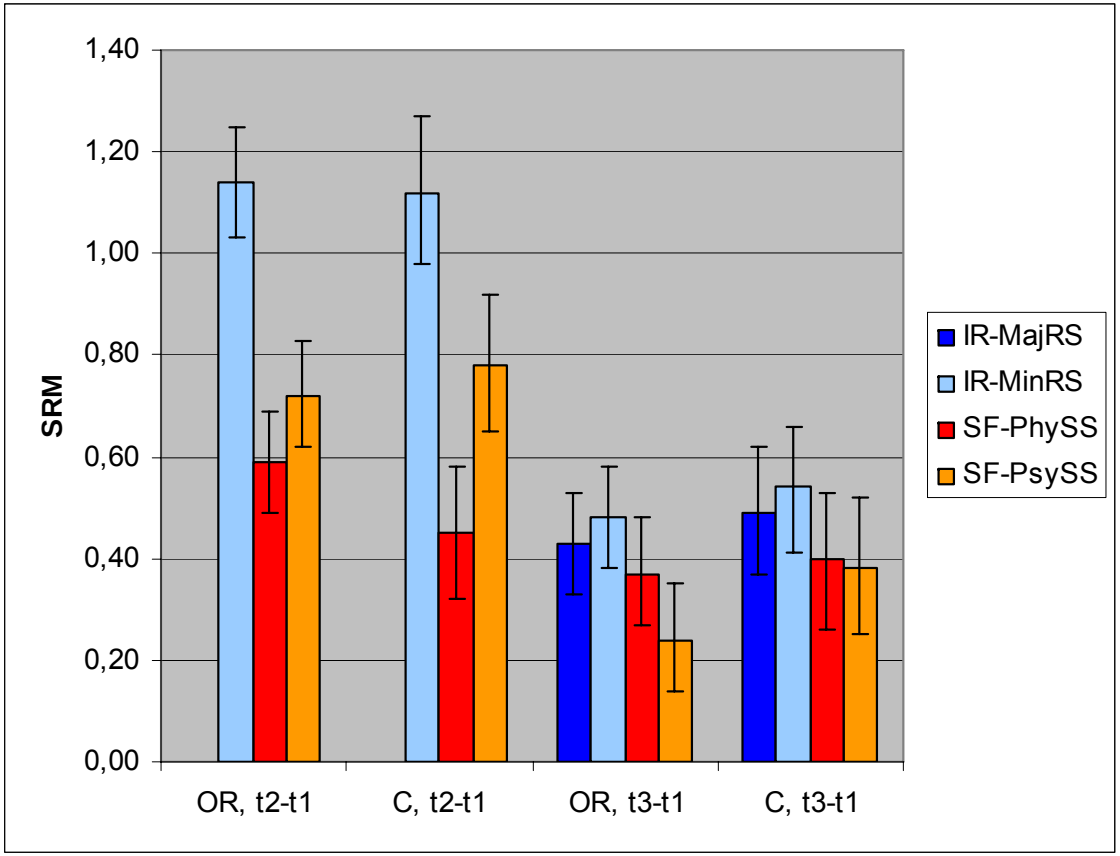


Fig. 1: Sensitivity to change beyond chance of Scores of ”General Health” (SRM, 95%-CIs)

General Health (II)

Table 1: Relative Sensitivity to change (SRM differences) of ”General Health”scores

	Orthopedics/Rheumatology			Cardiology		
SCORES	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**
Admission to Discharge						
MinRS – PhysHea	1.14 – 0.59	**0.55**** [0.43;0.72]	0.58	1.12 – 0.45	**0.69**** [0.43;0.91]	0.67
KleinRS – MentHea	1.14 – 0.72	**0.42**** [0.29;0.59]	0.44	1.12 – 0.78	**0.34**** [0.16;0.61]	0.38
Admission to Follow-Up						
MajRS – PhysSS	0.49 – 0.38	**0.06 (ns)** [-0.07;0.19]	0.06	0.49 – 0.40	**0.09 (ns)** [-0.05;0.27]	0.11
MajRS – MentHea	0.43 – 0.24	**0.19**** [0.10;0.37]	0.23	0.49 – 0.38	**0.11 (ns)** [-0.02;0.31]	0.15

Note: ns = not significant, 99%-CIs in squared brackets

Somatic Health/Pain (I)

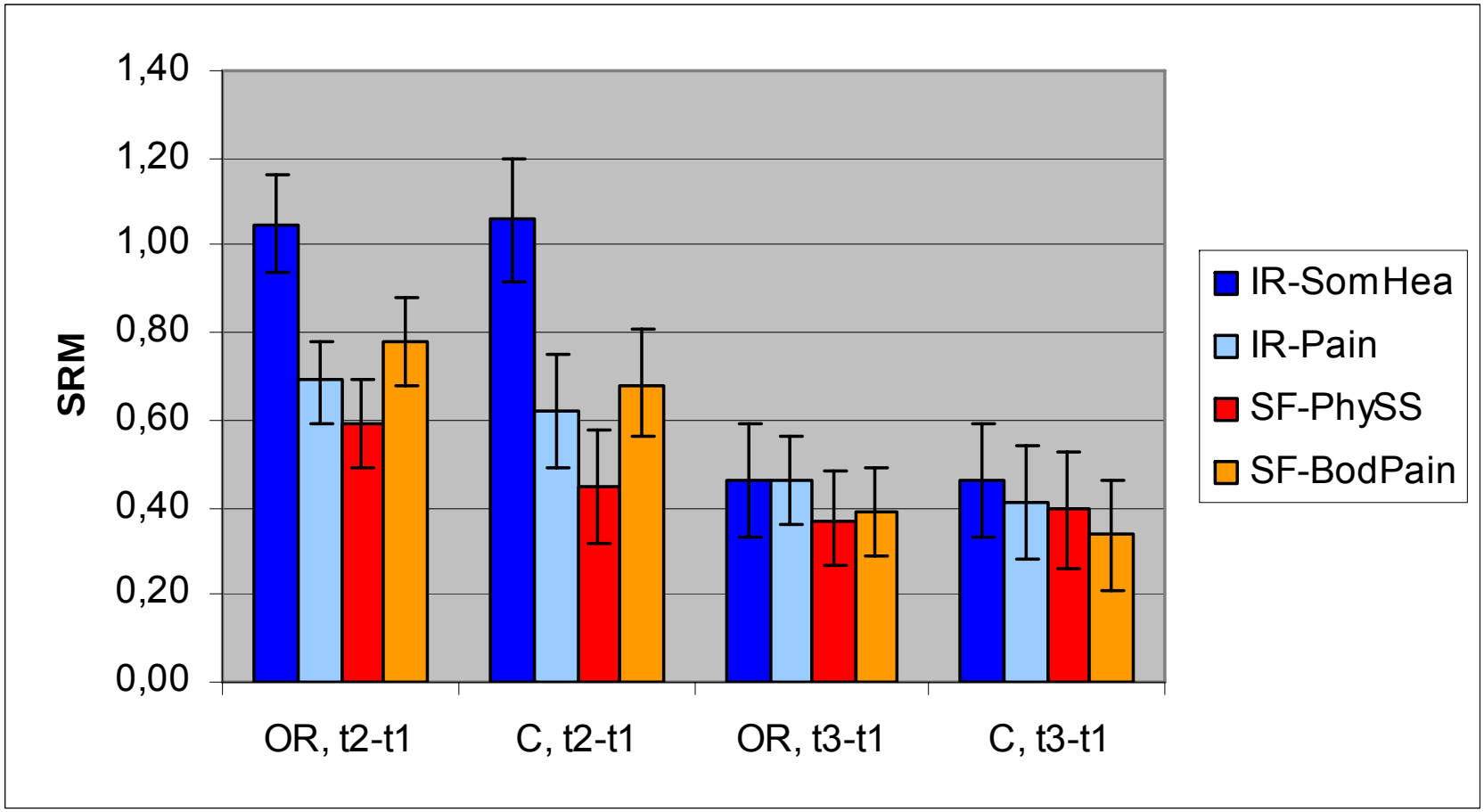


Fig. 2: Sensitivity to change beyond error of scores indicating ”Somatic Health”/”Pain” (SRM, 95%-CIs)

Physical Health/Pain (II)

Table 2: Relative sensitivity to change (SRM differences) of scores measuring ”Physical Health” / ”Pain”

	Orthopedics/Rheumatology			Cardiology		
SCORES	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**
Admission to Discharge						
SomHea – PhyHea	1.05 – 0.59	**0.46**** [0.33;0.62]	0.47	1.06 – 0.45	**0.61**** [0.39;0.83]	0.60
Pain – BodPain	0.69 – 0.78	**- 0.09 (ns)** [-0.22;0.01]	-0.11	0.62 – 0.68	**- 0.06 (ns)** [-0.25;0.08]	-0.08
Admission to Follow-Up						
SomHea – PhyHea	0.46 – 0.37	**0.09 (ns)** [-0.04;0.21]	0.09	0.46 - 0.40	**0.06 (ns)** [-0.07;0.26]	0.09
Pain– BodPain	0.46 – 0.39	**0.07 (ns)** [-0.12;0.09]	-0.01	0.41 – 0.34	**0.07 (ns)** [-0.09;0.21]	0.06

Note: ns = not significant, 99%-CIs in squared brackets

Mental Health (I)

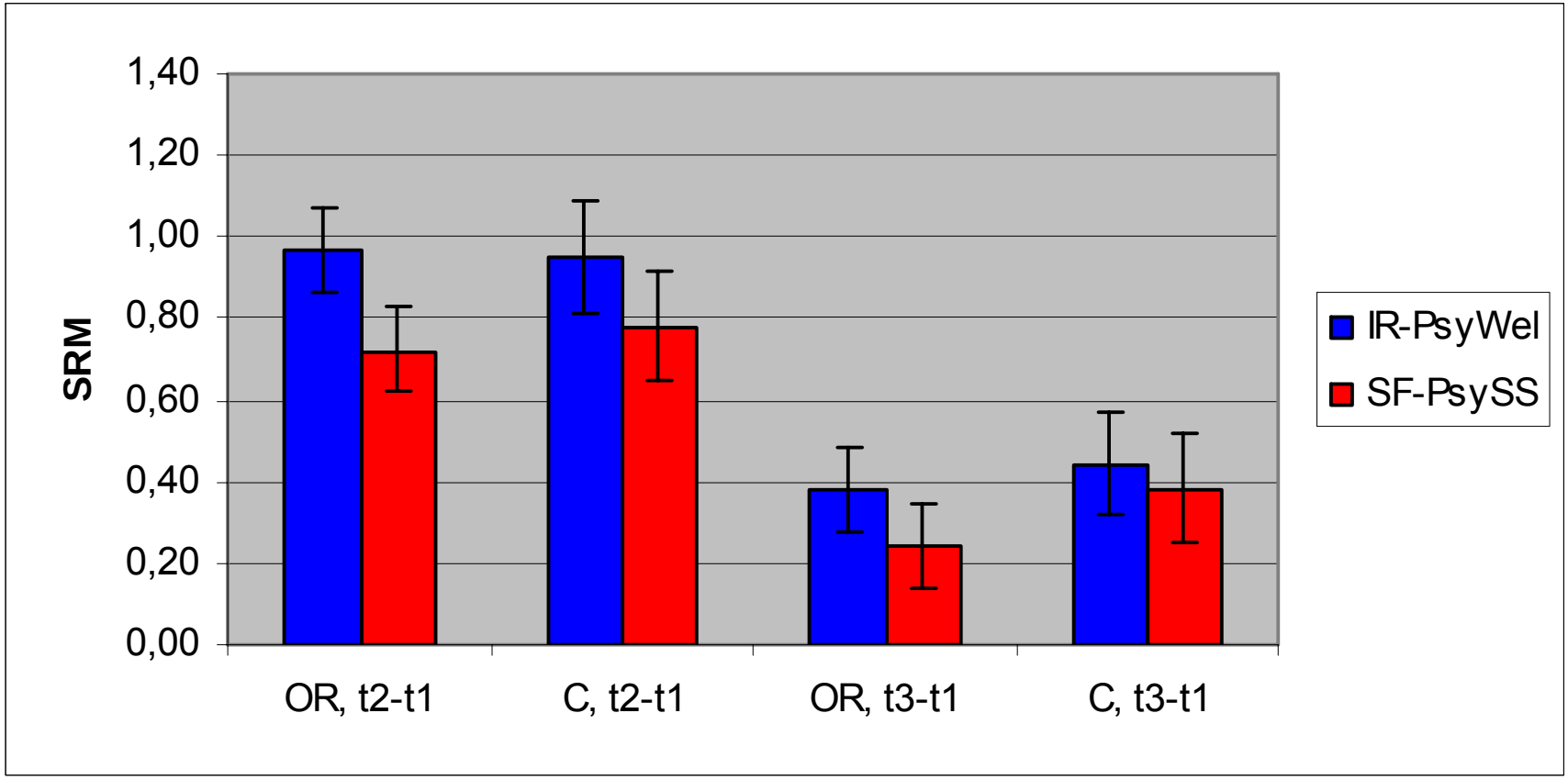


Fig. 3: Sensitivity to change beyond error of scores measuring ”Mental Health” (SRM, 95%-KIs)

Mental Health (II)

Table 3: Relative sensitivity to change (SRM differences) of scores measuring ”Mental Health”

	Orthopedics/Rheumatology			Cardiology		
SCORES	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**	**IRES3 – SF36 (SRM)**	**DIFF (SRM)**	**$DIFF_{Res}$**
Admission to Discharge						
PsyWel – MentHea	0.97 – 0.72	**0.25**** [0.15;0.42]	0.28	0.95 – 0.78	**0.17**** [0.04;0.40]	0.21
Admission to Follow-Up						
PsyWel – MentHea	0.38 – 0.24	**0.14**** [0.07;0.30]	0.19	0.44 – 0.38	**0.06 (ns)** [-0.06;0.26]	0.10

Note: ns = not significant, 99%-CIs in squared brackets

III. Summary and Conclusion

Sensitivity to Change and Sample Size

	Rehabilitation Admission to Discharge (3 Wks.)		Rehabilitation Admission to Follow-Up (3 Mts.)	
Construct	**Orthopedics/ Rheumatology**	**Cardiology**	**Orthopedics/ Rheumatology**	**Cardiology**
General Health	IRES-3 (9:18,9:25)	IRES-3 (9:15, 9:41)	IRES-3 (45:139, 45:60)	IRES-3 (35:60, 35:52)
Physical Health	IRES-3 (10:25)	IRES-3 (10:40)	IRES-3, SF-36 (40:54)	IRES-3, SF-36 (49:70)
Pain	SF-36, IRES-3 (15:19)	SF-36, IRES-3 (19:23)	IRES-3, SF-36 (40:54)	IRES-3, SF-36 (49:70)
Mental Health	IRES-3 (11:18)	IRES-3 (11:15)	IRES-3 (57:139)	IRES-3, SF-36 (43:57)

Note: Optimal sample sizes based on t-tests for dependent samples (α=0.05, β=0.20, Power=0.80, and observed standardized effect sizes, SRM)

Limitations

- Representativeness of the sample (e.g. inclusion criteria, drop-out rate)
- Representativeness of interventions and measurement points
- Bias due to effects of sequence and selective response patterns
- Comparability of questionnaire scores
- Suitability of descriptive and inferential statistical methods
  (e.g. type I error, coefficients, tests…)
- Practical aspects (e. g. length, redundancy)

Conclusion

- Sensitivity to change is an important concept in measurement theory.
- This criterion is especially relevant for intervention studies.

- Measurement instruments have to be compared in head-to-head studies, similar to intervention studies (RCTs).

- Generally speaking, the IRES-3 scores have equal or higher sensitivity to change than comparable scores of the SF-36 or the SCL-90-R.

- The observed differences in sensitivity to change can have dramatic consequences on optimal sample sizes in intervention studies.

Thank you!

Contact: wilmar.igl@ki.se

References

- Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist 12, 671-684.
- Aiken, L. R. (2000). Psychological Testing and Assessment (10th ed.). Boston: Allyn and Bacon
- Rost, J. (2004). Lehrbuch der Testtheorie – Testkonstruktion [Textbook of test theory – test construction] (2. Aufl.). Bern: Huber.
- Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F. & Bossuyt, P. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Quality of Life Research 12, 349-362.
- Igl, W., Zwingmann, C. & Faller, H. (2006). Änderungssensivitität von Fragebogen zur Erfassung der subjektiven Gesundheit – Ergebnisse einer prospektiven vergleichenden Studie [Sensitivity to change of questionnaires measuring subjective health- results of a prospective comparative study]. Rehabilitation 45(4), 232-242.
- Igl, W. (2007). Änderungssensitivität und Responsivität von generischen Patientenfragebogen in der Rehabilitation [Sensitivity to change and responsiveness of generic patient questionnaires in rehabilitation]. Dissertation, Department of Psychology, Albert-Ludwigs-Universität Freiburg. Published at http://www.freidok.uni-freiburg.de/volltexte/3015 on June 4, 2007.
- Bührlen, B., Gerdes, N. & Jäckel, W. H. (2005). Entwicklung und psychometrische Testung eines Patientenfragebogens für die medizinische Rehabilitation (IRES-3) [Development and psychometric testing of a patient questionnaire for medical rehabilitation (IRES-3), German Version]. Rehabilitation 44(2), 63-74.
- Bullinger, M. & Kirchberger, I. (1998). SF-36 Fragebogen zum Gesundheitszustand – Handanweisung [SF-36 Health Survey – Manual, German Version]. Göttingen: Hogrefe.
- Franke, G. H. (2002). SCL-90-R Symptom-Checkliste von L. R. Derogatis – Deutsche Version [SCL-90-R Symptom-Checklist by L. R. Derogatis – German Version] (2. Aufl.). Göttingen: Beltz.