Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet

Stockholm, October 31, 2006

Questionnaires in Medical Research

- Theory and Evaluation Methods -

Wilmar Igl

Clinical Trial Center,
University Hospital of Würzburg

I. Introduction

Questionnaires as Measuring Instruments in Medicine

- Evidence-based medicine requires the evaluation of the effectiveness of medical treatments
- Two categories of evaluation measures:
  - biometric measures, e. g. blood pressure, joint play
  - psychometric measures, e. g. depression, subjective health
- Psychometric measures are important, in particular regarding patients with chronic diseases (e.g. cardiac insufficiency, diabetes, multiple sclerosis)
- The quality of measuring instruments has a strong effect on study design (e.g. sample size) and outcomes (e.g. effect size, interpretation)
- Problems to select an adequate questionnaire and apply it properly in medical research projects

Overview

I. Introduction

II. Theory

1. What‘s a Questionnaire?
2. Classical Test Theory (CTT)

III. Evaluation Methods

1. Reliability
2. Validity
3. Sensitivity to Change

IV. Conclusion

II. Theory

1. What‘s a Questionnaire?

What's a Questionnaire?

- Definition:
  A questionnaire is an ordered list of standardized questions and answers for systematic data collection on one or more individuals of a population providing an aggregated score indicating the degree of the construct of interest.

- Possible topics:
  behavior (e.g. eating habits), emotions (e.g. pain), cognition (e.g. health belief model), life events (e.g. heart attack) and many more

- In medical research the concept "Subjective Health" (cf. HRQOL) is most relevant:
  - generic: SF-36, NHP, WHO-QOL-100, SIP, SCL-90-R
  - specific: SMFA, WOMAC, KCCQ, QLQ-C30

Application of Questionnaires as...

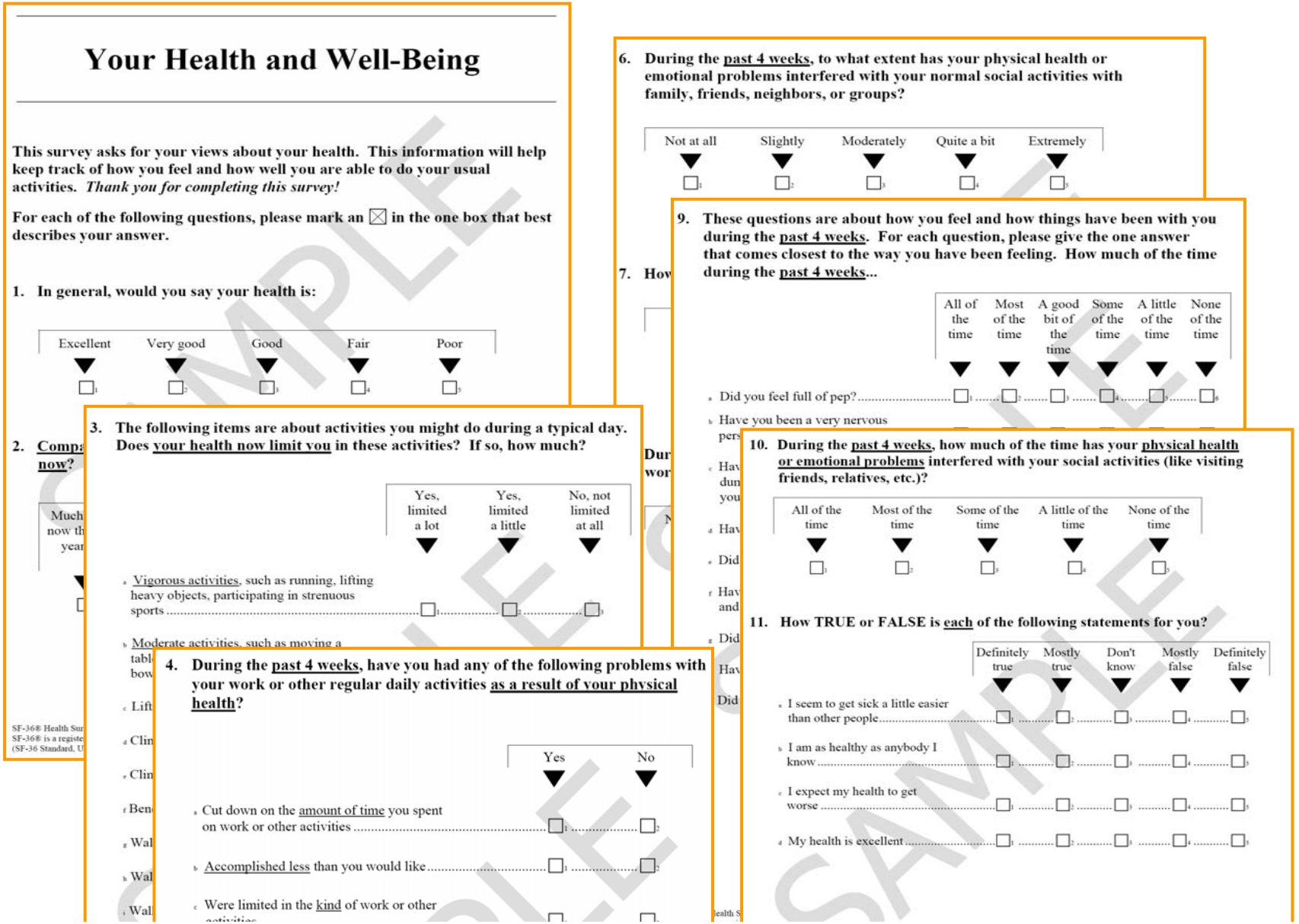
Diagnostic test

- Context: clinical practice, diagnosis
- Outcome: (quantitative) score of the relative degree of a stabile trait of one individual (compared to a functional population)
- Analysis: descriptive, individual multi-dimensional profile
- Test criteria: reliability and validity on person level, norm values (!)

Research tool

- Context: research, discrimination, evaluation or prediction
- Outcome: quantitative score of the degree of a variable state in a group of individuals (compared to another dysfunctional group)
- Analysis: descriptive, inferential, single, aggregated score of a group
- Test criteria: reliability and validity on group level

SF-36 (Ware et al., 1993)



Your Health and Well-Being

This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. ***Thank you for completing this survey!***

For each of the following questions, please mark an ☒ in the one box that best describes your answer.

1. In general, would you say your health is:

Excellent | Very good | Good | Fair | Poor

2. Compa… now?

Much now th… year

3. The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports			
b. Moderate activities, such as moving a tabl… bow…			
c. Lift…			
d. Clim…			
e. Clim…			
f. Bend…			
g. Wal…			
h. Wal…			
i. Wal…			

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	Yes	No
a. Cut down on the amount of time you spent on work or other activities		
b. Accomplished less than you would like		
c. Were limited in the kind of work or other activities		

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

Not at all | Slightly | Moderately | Quite a bit | Extremely

7. How…

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks...

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
a. Did you feel full of pep?						
b. Have you been a very nervous pers…						
c. Hav… dun… you…						
d. Hav…						
e. Did…						
f. Hav… and…						
g. Did…						
Hav…						
Did…						

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

All of the time | Most of the time | Some of the time | A little of the time | None of the time

11. How TRUE or FALSE is each of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a. I seem to get sick a little easier than other people					
b. I am as healthy as anybody I know					
c. I expect my health to get worse					
d. My health is excellent					

SF-36® Health Sur… SF-36® is a registe… (SF-36 Standard, U…

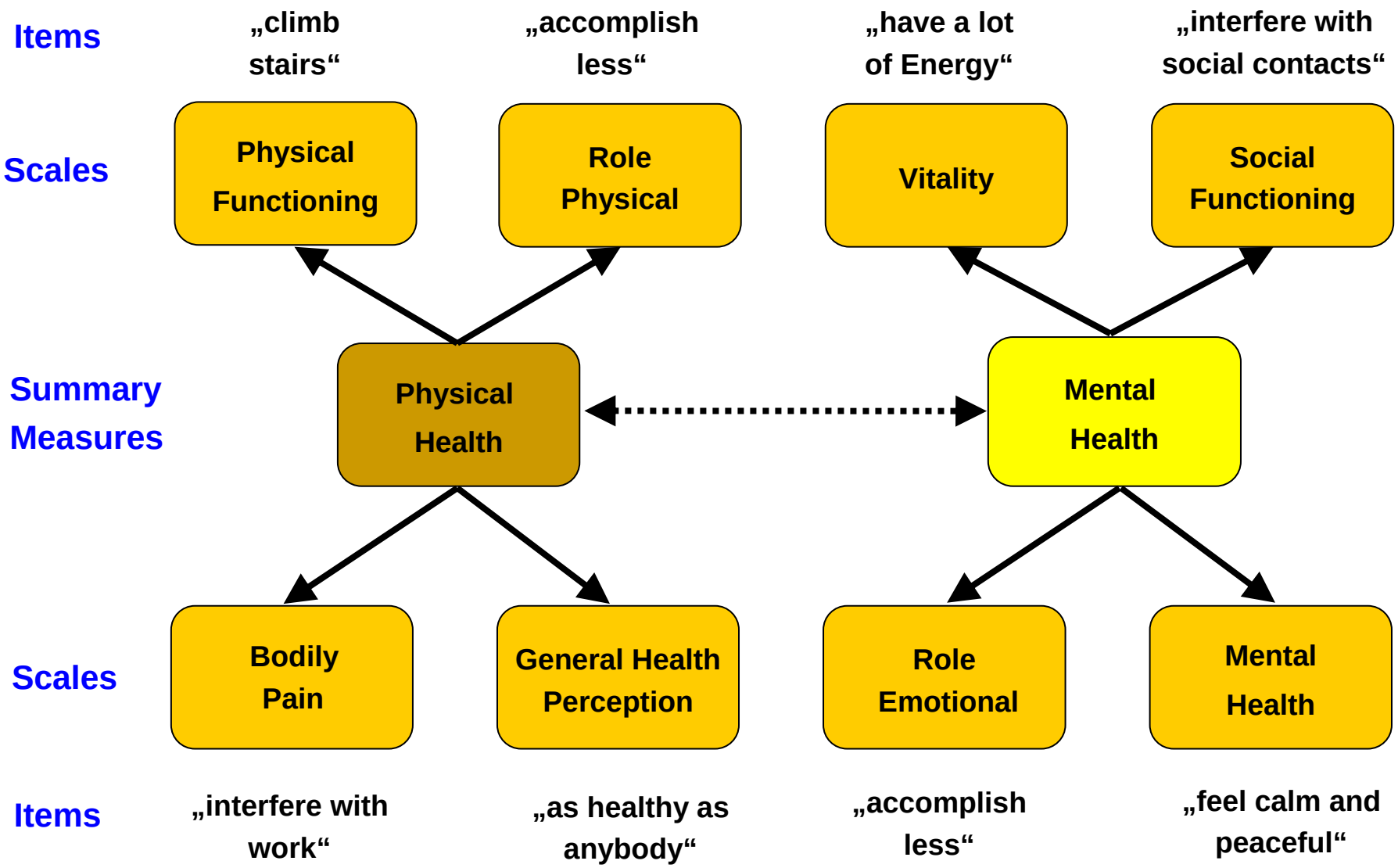
SAMPLE

URL: www.sf-36.org/tools/pdf/SF-36v1_Standard_Sample.pdf

SF-36 ... in a Nutshell

- Construct: subjective health (generic)
- Administration: patient, physician, interviewer
- Recall versions: 1 week, 4 weeks
- Length: 36 items
- Response options: multiple choice (2 to 6)
- Scoring: 2 summary measures, 8 scales

SF-36: Measurement Model



2. Classical Test Theory

Test Theory in General

- Test theories aim at defining conditions which have to be fulfilled to be able to infere the degree of abstract constructs from observed item reponses.

Example:

- Construct: "health-related quality of life"
- Question: „How much difficulty do you have when walking fast?"
- Response: ❑ none ❑ some ❑ much ☑ very much

Classical Test Theory

- Classical statistical model of measurement applied in physics serves as a basic paradigm.

- Deterministic approach (vs. probabilistic approach, IRT theories)

- Basic equation:
  The test result is the sum of a true value (T) and an unsystematic, random measurement error (E) of repeated measurements on a single individual.

$$X_1 = T_1 + E_1$$

- Classical test theory can be applied to all sorts of measurements,
  i. e. tests measuring stabile traits or variable states.

„Axioms“ of Classical Test Theory

1. $\mu (E_1) = 0$ or $\mu(X_1) = T_1$
   Errors of repeated measurements add up to zero.
2. $\rho (T_1, E_1) = 0$
   Measurement error and the true value of the interesting trait are not correlated.
3. $\rho (E_1, T_2) = 0$
   Measurement error and the true value of other constructs are not correlated.
4. $\rho (E_1, E_2) = 0$
   Measurement errors of different traits are not correlated.

Independence of measurement errors

Major Test Criteria (I)

- The quality of a test can be defined by its variance components.

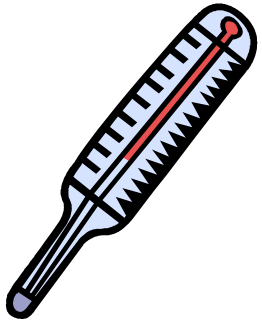
- Reliability: Degree of precision of measuring a construct without considering whether this is the construct of interest

$$\text{Rel}(X_1) = \frac{\sigma^2(T_1)}{\sigma^2(X_1)} = \frac{\sigma^2(T_1)}{\sigma^2(T_1) + \sigma^2(E_1)}$$

- Validity: Degree of accuracy of measuring the construct of interest

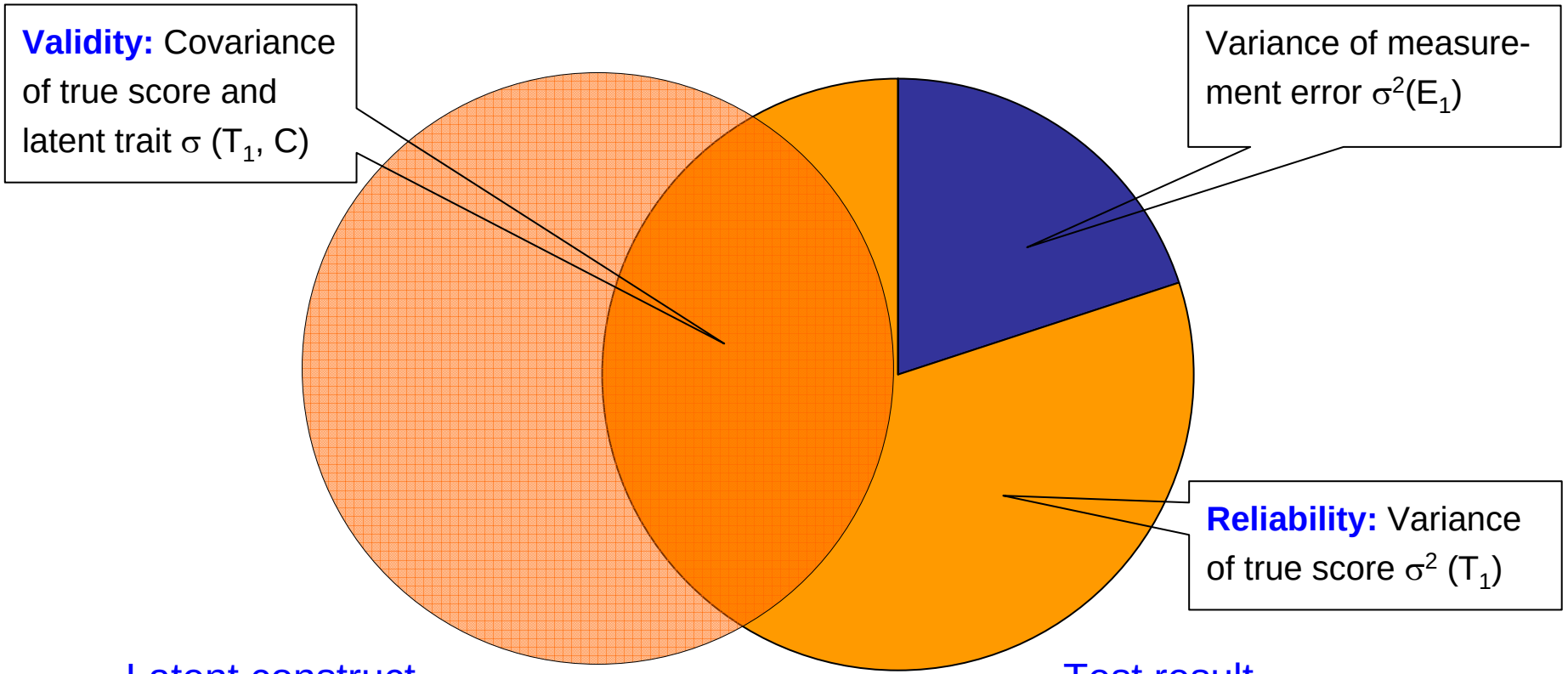
$$\text{Val}(X_1) = \frac{\sigma(X_1, C)}{\sigma(X_1) \cdot \sigma(C)}$$

- Example: Measuring „intelligence“ with a clinical thermometer



Major Test Criteria (II)

Variance components of measurement error, true score and latent trait (criterion)



Minor Test Criteria

- Standardization to compare a person's test results relative to norm values of other individuals of the same population
- Comparability with other instruments measuring the same construct.
- Popularity of the instrument among researchers/reviewers or clinicians.
- Practicability to minimize expenses, e.g. on time, costs, material.
- Acceptance among patients, e.g. non-applicable questions or items eliciting embarrassment may reduce it.

Standardization

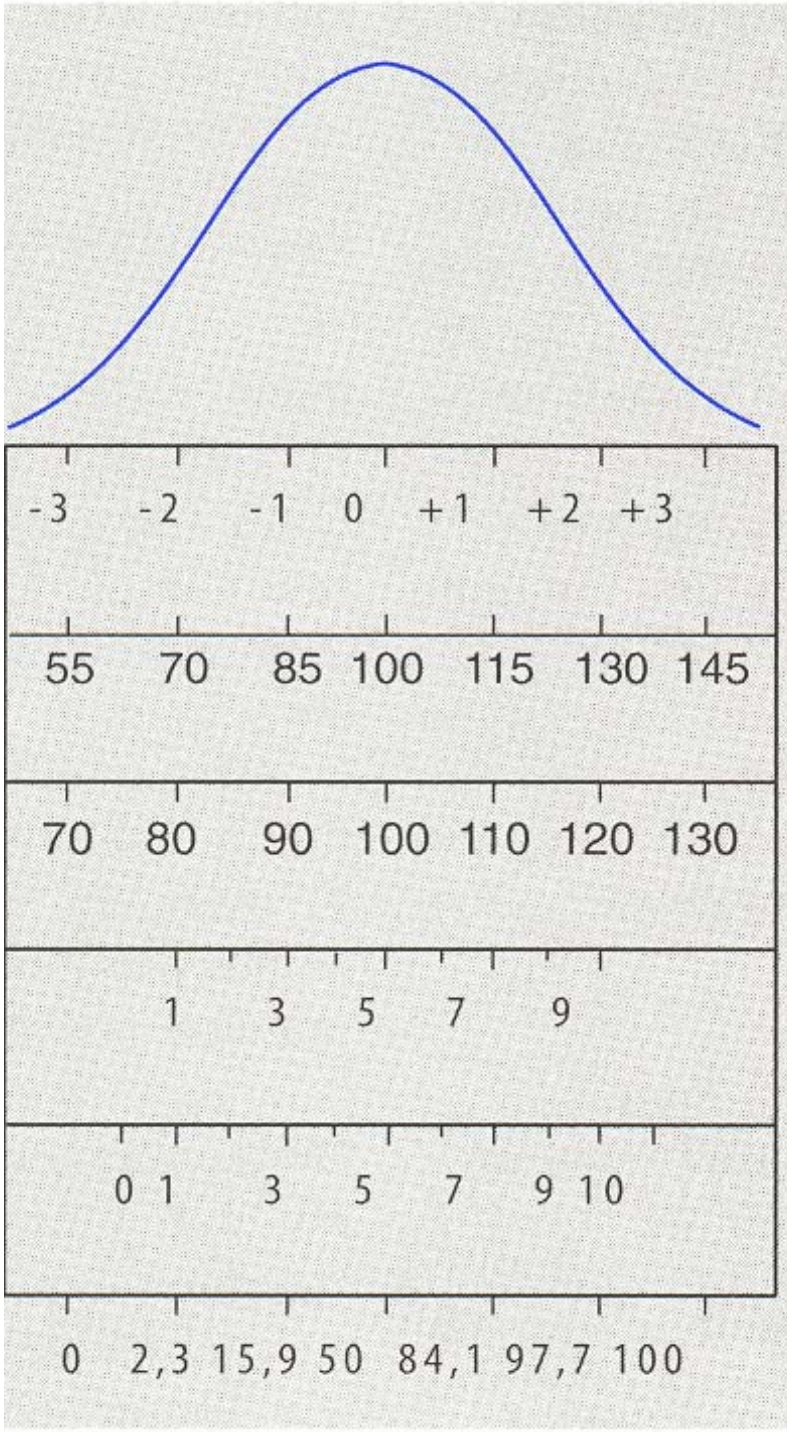
- Procedure:
  - collect test values in a standardization sample representative of the overall population
  - transform raw scores to standardized score norms:
    - IQ scores: $\mu = 100$, $\sigma = 15$
    - Z scores: $\mu = 100$, $\sigma = 10$
    - T scores: $\mu = 50$, $\sigma = 10$
    - Stanine scores: $\mu = 5$, $\sigma = 2$
    - Percentile score: [0;100%]

General equation :

$$X_T = \frac{\sigma}{s} \cdot X + \left(\mu - \frac{\sigma}{s} \overline{X} \right)$$

- Score transformations allow the comparison of an individual to a representative sample of the same population independent of test characteristics, e.g. length, difficulty, raw score values.

Standard score norm		Transformation equation
Standard normal distribution	-3 -2 -1 0 +1 +2 +3	$z = (X - M_X) / SD_X$
IQ scores	55 70 85 100 115 130 145	IQ = 100 + 15 z
Z scores	70 80 90 100 110 120 130	Z = 100 + 10 z
Stanine scores	1 3 5 7 9	St = 5 + 2 z
Sten scores	0 1 3 5 7 9 10	C = 5 + 2 z
Percentile scores	0 2,3 15,9 50 84,1 97,7 100	non-linear



Lang & Faller 1998, p. 16

III. Evaluation Methods for Test Criteria

1. Reliability

Reliability

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

- Synonyms: precision
- Definition: Fraction of variance of true score variance compared to total variance of observed test scores.
- Basic approach: Estimation of variance fractions by correlating two (or more) measurements of a test which are supposed to be equal.
- Coefficients: Pearson correlation (r), intraclass correlation coefficient (ICC), concordance correlation coefficient (CCC)
- Valuation rule (of thumb):
  - „high“: Rel = [0.90; 1.00]
  - „medium“: Rel = [0.80; 0.90]
  - „low“: Rel = [0.70; 0.80]
  - „insufficient“: Rel < 0.70

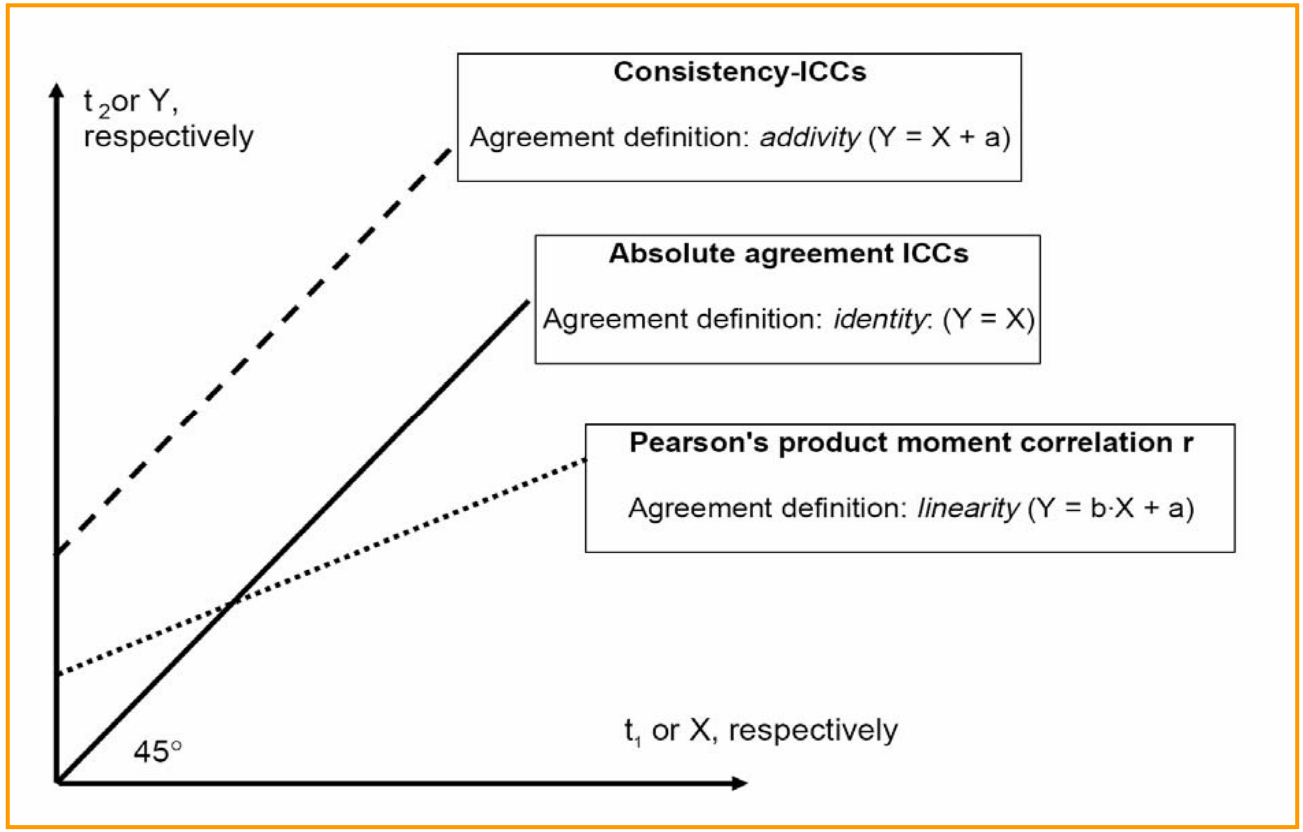
Some Notes on Reliability Coefficients (I)

- **Pearson Correlation Coefficient r** (Pearson, 1907)
  - agreement definition: linearity
  - no. of measurements: two (e.g. repeated measures)
  - easy to compute, most widely applied

- **Intraclass Correlation Coefficient ICC** (Shrout & Fleiss, 1979)
  - agreement definition: additivity vs. identity
  - no. of measurements: two or more (e.g. repeated measures)
  - requires equal variances
  - calculation more complex, popularity increasing

- **Concordance Correlation Coefficient CCC** (Lin, 1989)
  - agreement definition: identity
  - no. of measurements: two (e.g. repeated measures)
  - differences of means and variability reduce reliability
  - easy to calculate, hardly applied

Some Notes on Reliability Coefficients (II)



Comparison of various correlation coefficients
as measures of reproducibility/test–retest reliability (Schuck, 2004).

Methods of Reliability Estimation

- Test-retest method:
  agreement of measurements repeated over time

- Parallel-forms method:
  agreement of measurements of "equivalent" tests

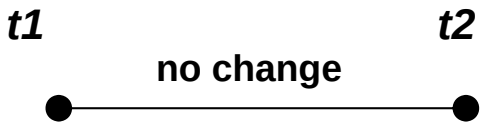
- Split-half method:
  agreement of two halves of a test

- Internal consistency:
  agreement of items within a (homogenous) scale

Test-Retest Method

$$Rel_{Retest} = r_{t_1,t_2} = \frac{\sigma(t_1, t_2)}{\sigma(t_1) \cdot \sigma(t_2)}$$

- Synonyms: stability, reproducibility
- Study design: repeated application of the test to a sample in an adequate time interval



- Assumptions:
  - stability of the trait over time
  - no effects of sequence, e.g. recall of former responses
- Problems:
  - overestimation of reliability if time interval is too short, e.g. because of memory effects
  - underestimation of reliability if time interval is too long, e.g. due to true changes
  - expenses on time and effort (drop out!)

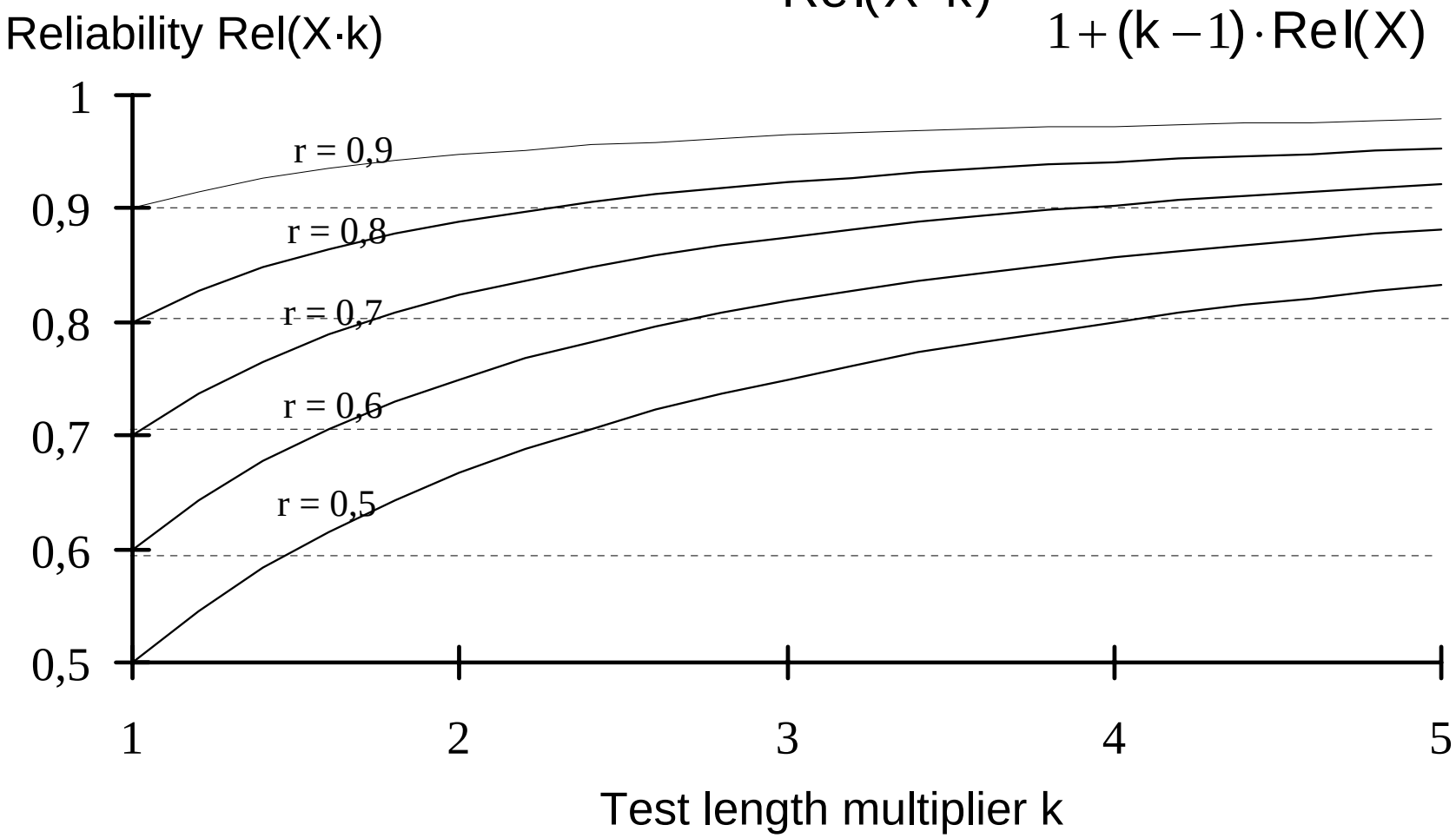
Internal Consistency

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} s^2_{item}}{s^2_{test}}\right)$$

- Study design: single application of the test to a sample
- Method: Partitioning in as many „subtests“ as items, extending the concept of split-half reliability
- Coefficients:
  - Formulas by Kuder & Richardson (1939)
  - Coefficient α by Cronbach (1951)
- Cronbachs α
  - Interpretation: average split-half reliability (Pearson correlation) over all possible test halves (~ homogeneity index)
  - Requirements: polytomous items (cf. KR20 for dichotomous items)
- Pros: minimal expenses
- Cons: Underestimation of reliability of heterogenous and multi-dimensional scales

Test Length and Reliability

$$Rel(X \cdot k) = \frac{k \cdot Rel(X)}{1 + (k - 1) \cdot Rel(X)}$$



Rost 2004, p. 357

2. Validity

Validity

- Synonyms: *unbiasedness*, *accuracy*
- Definition: fraction of shared variance of test result and the trait of interest
- Coefficients: Pearson correlation r, η^2 (eta-square), contingency coefficient C
- Valuation rule (of thumb):
  - „high“: $r_{tc} > 0.60$
  - „medium“: $r_{tc} = [0.40; 0.60]$
  - „low“: $r_{tc} < 0.40$
  - „significant“: $r_{tc} >> 0$ (H_0: $r_{tc} = 0$, $p < 0.05$)

Content Validity

- Synonyms: (face validity), logical validity

- The content of a test represents the essential aspects of the domain of interest (in the opinion of experts).

- Examples:
  - *Typing skill*, e.g. simple sensoric and motoric tests
  - *Numerical intelligence*, e.g. basic arithmetic operations
  - *Physical well-being*, e.g. „Do you have pain in your body?“

- Possible quantitative indicators of content validity,
  e.g. fraction of missing data, floor and ceiling effects

- Problems:
  - Difficult to quantify
  - Lack of aggreement of experts

Criterion Validity

$$\text{Val} = r_{tc} = \frac{\sigma(T,C)}{\sigma_T \cdot \sigma_C}$$

- Consistency of the test score with a valid external criterion of the construct of interest
- Example: Academic aptitude test and grade of the final exam.

- Coefficient: Pearson correlation between test and criterion, Area under curve (ROC analysis)

- Problems:
  - An adequate criterion ("gold standard") is not available, e.g. measuring religiousness by measuring the frequency of attendance at church
  - Interpretation of validity coefficients is difficult if the external criterion is unreliable or invalid.

Types of Criterion Validity

- Concurrent validity: test and criterion are measured at the same time

- Prognostic/predictive validity: test result is measured before criterion
  Example: psychoticism ⇨ psychiatric diagnosis

- Discriminative validity: definition of the criterion by the affiliation of subjects to a distinct group (patients vs. non-patients)

Construct Validity

- Examination of a system of hypotheses, which can be derived from the construct of interest

- Example: „Subjective Health“
  Scales measuring physical health of test A are supposed to show higher correlations with other somatic scores of test B than with its scales for mental health.

- Multitrait-multimethod approach (Campbell & Fiske, 1959):
  sophisticated approach for testing a set of assumptions based on the concepts of convergent and discriminant validity

- Problems:
  - valid criteria are necessary
  - hypotheses have to be confirmed
  - confounding of correct specification of hypotheses and validity of the criteria

Other Types of Validity

- Differential validity: validity of the test result in different populations (also: discriminative validity)
- Factorial validity: correlation (factor loading) of item and the relevant latent factor (cf. factor analysis)
- External validity: correlation with a valid, external criterion
- Internal validity: test model is valid for the data (item or person score)
- Longitudinal validity: correlation with true change
- Warning! Definition of validity in experimental trials!
  - external validity: generalizability of results to a certain population of individuals
  - internal validity: causal relationship between the independent variable (treatment) and the dependent variable (outcome)

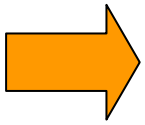
3. Sensitivity to Change

Background

- Applications of clinical measurement instruments (Kirshner & Guyatt, 1985):
  - Discrimination: measuring stable constructs, e.g. intelligence
  - Prediction: forecasting of values/events, e.g. mortality
  - Evaluation: measurement of change, e.g. treatment succes

- Tests for discrimination or prediction: conventional criteria and methods are sufficient

- Tests for evaluation: conventional criteria and methods are not adequate (e.g. test-retest reliability)



Measurement of sensitivity to change is essential!

Sensitivity to Change

- 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

Example (cf. SIP, Sickness Impact Profile):

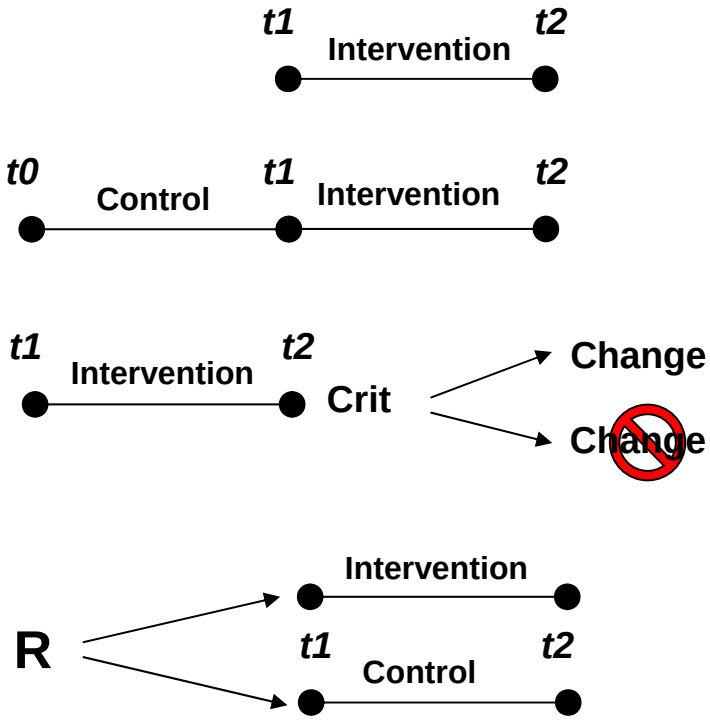
Question: "Have you ever attempted suicide?"

Answer: ☐ yes ☐ no

- Question is adequate for differentiating between groups with different emotional stress (discrimination)
- Question is inadequate for measuring change caused by a clinical intervention (evaluation)

Study Designs

- Designs:
  - single-group designs
    - simple pre-post design
    - pre-post design with baseline
    - simple pre-post design with external criteria
  - two-group designs (with experimental and control group)

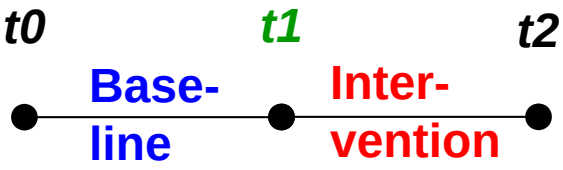


Coefficients

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

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

Standardized Effect Sizes



$$SES = \frac{M(t2) - M(t1)}{SD(t1)}$$

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

$$SRM = \frac{M(t2) - M(t1)}{SD(t2 - t1)}$$

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

$$GRI = \frac{M(t2) - M(t1)}{SD(t1 - t0)}$$

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

IV. Conclusion

Some practical considerations

- Reliability measures
  - depend strongly on variability of individuals
  - Example: health-related quality of life questionnaire
    Rel(overall population) > Rel(young men)

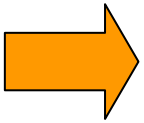
- Validity measures
  - depend strongly on variability of individuals
  - "gold standards" are hard to find
    Example: single criterion for subjective health, e.g. physical performance?

- Measures of sensitivity to change
  - depend strongly on the effect of the intervention
  - difficult to compare across studies

- Head-to-head studies comparing questionnaires under similar conditions are most valid.

Summary

- Questionnaires are versatile measuring instruments in medical research
- Evaluation of the quality of questionnaires on the basis of (classical) test theory (reliability, validity, sensitivity to change) is essential
- Various methods available:
  - Reliability (test-retest, internal consistency,...)
  - Validity (content v., criterion v., construct validity,...)
  - Sensitivity to change (one sample vs. two sample designs, distribution-based vs. criterion-based coefficients)
- Selection of methods depends on the application of the instrument (discrimination, prediction, evaluation)
- Test criteria are not a constant feature of the instrument, but depend strongly on the specific context of a trial (sample, intervention, measurement points...)



Evaluation is not a single event, but a process.

Thank you!

Contact: wilmar.igl@uni-wuerzburg.de

Würzburg's Old Main Bridge (1473–1543) and Fortress Marienberg (~1600)

References

- Aiken, L. R. (2000). Psychological Testing and Assessment (10th ed.). Boston: Allyn and Bacon
- Anastasi, A. & Urbina, S. (1997). Psychological Testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Bortz, J. (2002). Forschungsmethoden und Evaluation (3rd ed.). Berlin: Springer
- Kraemer, H. C. (1992). Evaluating medical tests. Newbury Park: Sage.
- Lord, F. M. & Novick, M. R. (1976). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Lienert, G. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: PVU.
- Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion. Bern: Huber.
- Steyer, R. & Eid, M. (2001). Messen und Testen (2nd ed.). Berlin: Springer.