AARHUS UNIVERSITY

Validity

Advanced course in questionnaire technique and clinimetrics

Eva Ørnbøl

1

AARHUS UNIVERSITY

Outline

- Secrets to measuring a piece of paper
- The concept of validity – and why it depends on the context
- Content and face validity
- Criterion validity
  - Concurrent and predictive validity
  - Dichotomous, ordinal and continuous instruments
- Construct validity
  - Structural validity
  - Hypotheses testing: direction, magnitude and the 75% criterion
  - Correlations
  - Cross-cultural validity
- Validation in context: sample size and missing data

2

AARHUS UNIVERSITY

Validity: introductory course vs. today

From the introductory course

- COSMIN taxonomy and definitions
- Content validity: relevance, comprehensiveness, comprehensibility
- Steps in hypothesis testing
- Cross-cultural validity: definition
- Bland-Altman limits of agreement (as measurement error)

New today

- Criterion validity in practice: choosing the statistic by level of measurement
- Sensitivity, specificity, PPV/NPV and ROC/AUC – and how prevalence matters
- Bland-Altman for criterion validity
- Correlations: Pearson vs. Spearman, and their pitfalls
- Validation in context: sample size and missing data

AARHUS UNIVERSITY

Secrets to measuring a piece of paper

Measuring a piece of paper – how hard can it be?

A teacher gives an 8th-grade class a sheet of paper and a ruler, and asks them to calculate the area.

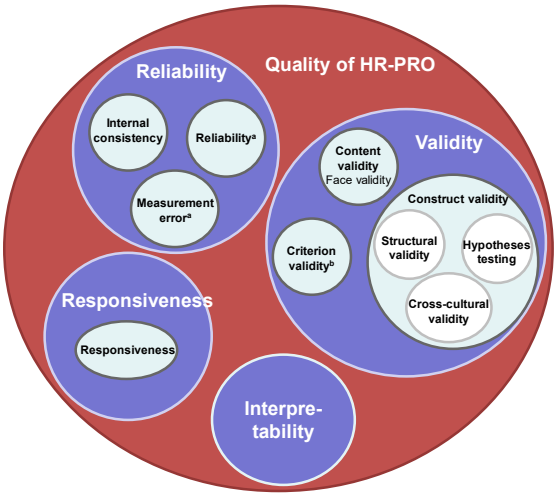
Easy pitfalls when measuring:

- taking the task at face value
- assuming the object is what it usually is
- trusting the instrument without checking it

https://www.youtube.com/watch?v=9yUZTTLpDtk

AARHUS UNIVERSITY

The COSMIN taxonomy



[a] test-retest, inter-rater, intra-rater
[b] concurrent validity, predictive validity

5

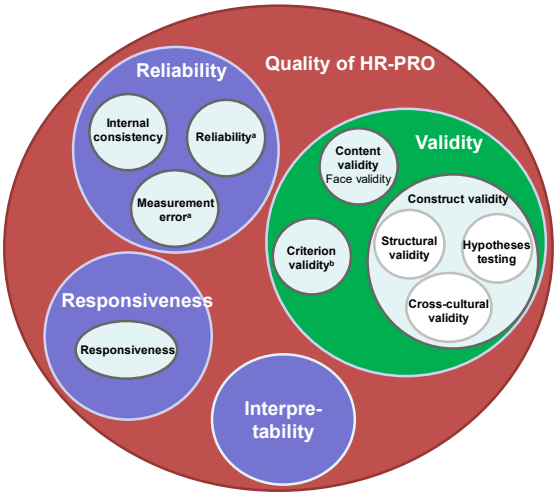
AARHUS UNIVERSITY

Validity

"The degree to which an instrument truly measures the construct(s) it purports to measure"

COSMIN (Mokkink et al. 2010); de Vet et al. 2011

- **Reliability:** do we get the same result when we measure again?
- **Validity:** do we measure the right thing?
- A measurement can be reliable without being valid – like the ruler in the video
- "Truly" cannot be tested directly – we test hypotheses about the scores



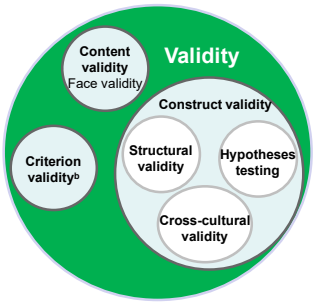
[a] test-retest, inter-rater, intra-rater
[b] concurrent validity, predictive validity

6

AARHUS UNIVERSITY

Validation: testing hypotheses

- Validation consists of testing hypotheses about
  - relationships with other constructs
  - differences between groups of patients
- If the hypotheses are confirmed, the scores are consistent with the theoretical model of the construct
- **Knowledge about the construct:** validity can only be assessed if the construct is clearly described (conceptual model)
- **Complexity of the construct:** a unidimensional construct is easier to validate than a multidimensional one – validate each dimension separately (e.g. fatigue)
- **Hypotheses:** as specific as possible, based on existing knowledge

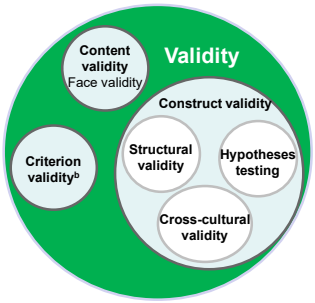


7

AARHUS UNIVERSITY

Validity depends on the context

- **Dependency on the situation:** validity depends on the population (e.g. age, sex, diagnosis) and the setting
  - reassess when the instrument is used in a new situation, e.g. a new language (cross-cultural validation)
- **We validate scores, not instruments:** an instrument is never "validated" in general
- **Validation is a continuous process:** knowledge about the construct grows during validation, and the hypotheses can become more precise



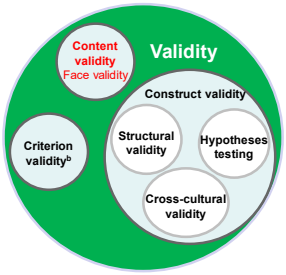
8

AARHUS UNIVERSITY

Content validity

"The degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured"

- Is the instrument relevant?
  - Do all items refer to relevant aspects of the construct?
  - Are all items relevant to the study population?
  - Are all items relevant to the purpose of the study?
- Is the instrument comprehensive?
  - Is the construct completely covered by the items?
- Is the instrument comprehensible?



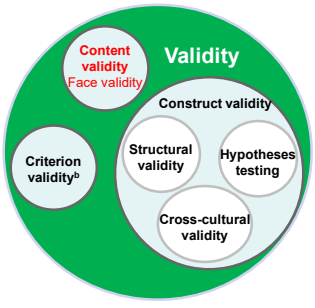
9

AARHUS UNIVERSITY

Face validity

"The degree to which a measurement instrument, indeed, looks as though it is an adequate reflection of the construct to be measured"

- A subjective assessment
- No standards with regard to assessment
- Lack of face validity → strong argument for not using an instrument

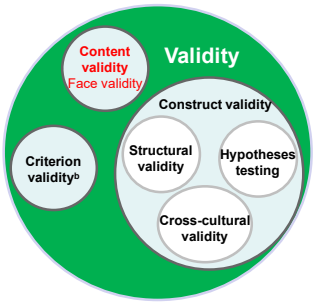


10

AARHUS UNIVERSITY

Content and face validity

- The process of content validation
  1. Consider information about construct and situation
  2. Consider information about content of the instrument
  3. Select an expert panel
  4. Assess whether content of the instrument corresponds with the construct (is it relevant, comprehensive and comprehensible)
  5. Use a strategy or framework to assess correspondence between the instrument and construct

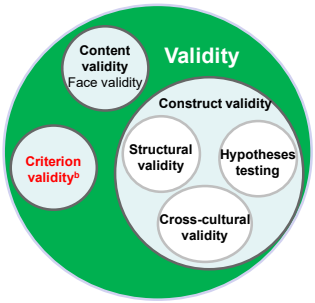


AARHUS UNIVERSITY

Criterion validity

"The degree to which the scores of an instrument are an adequate reflection of a gold standard"

- Can only be assessed when a gold standard is available
- What is a gold standard?
  - In theory: A perfectly valid assessment
  - Seldom exists in practice
  - In practice: An instrument which is regarded as ideal by experts
  - PRO's almost always lack a gold standard.

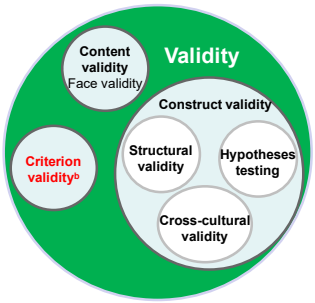


AARHUS UNIVERSITY

Criterion validity, subtypes

- Concurrent validity
  - Consider the score for the instrument and criterion at the same time
  - Usually assessed for instruments used for evaluative and diagnostic purposes
- Predictive validity
  - Consider whether the instrument predicts the criterion in the future
  - Used in predictive applications

The statistical methods on the following slides apply to both subtypes – the examples are concurrent, followed by a predictive example

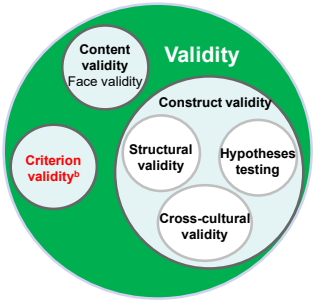


AARHUS UNIVERSITY

Criterion validity

Hypothesis: The instrument under study is as good as the gold standard!

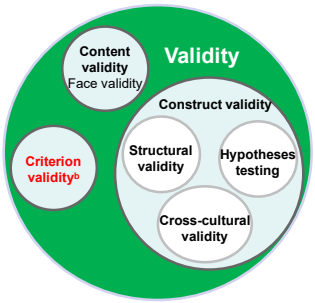
In practice the essential question is whether the instrument is sufficiently valid for its clinical purpose



AARHUS UNIVERSITY

Criterion validity

- The process of criterion validation
  1. Identify a suitable criterion and method of measurement
  2. Identify an appropriate sample of the target population
  3. Define a priori the required level of agreement between instrument and criterion
  4. Obtain scores for instrument and criterion independently
  5. Determine the strength of relationship between instrument scores and criterion scores (see next slide)



15

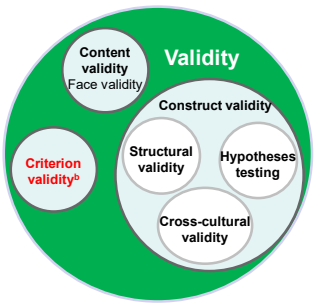
AARHUS UNIVERSITY

Criterion validity

Table 6.2: Overview of statistical parameters for various levels of meausrement for the gold standard and measurement instrument under study

Level of measurement		Same units	Statistical parameter
Gold standard	Measurement instrument		
Dichotomous	Dichotomous	Yes	Sensitivity and specificity
	Ordinal	NA	ROC
	Continuous	NA	ROC
Ordinal	Ordinal	Yes	Weighted kappa
		NO	Spearman's r
	Continuous	NA	ROC/Spearman's r
Continuous	Continuous	Yes	Bland-Altman/ICC
		No	Spearman's r or Pearson's r

H. C. W. de Vet, at al., Measurement in Medicine p. 163



16

AARHUS UNIVERSITY

Criterion validity, dichotomous instruments

Dichotomous gold standard

- Sensitivity
  - Sens= TP/(TP+FN)
  - Probability of a positive test given that the patient is truly sick
  - The true positive rate
  - Quantifies the avoiding of false negatives
- Specificity
  - Spec= TN/(FP+TN)
  - Probability of a negative test given that the patient is well
  - The true negative rate
  - Quantifies the avoiding of false positives

Table: Cross-tabulation

		Gold standard +	Gold standard -
Test result	+	TP	FP
	-	FN	TN

TP = true positive
FP = false positive
FN = false negative
TN = true negative

17

AARHUS UNIVERSITY

Criterion validity, dichotomous instruments

Dichotomous gold standard

- Sensitivity and specificity
  - Test specific: describe how well the test performs against the gold standard
  - A perfect test has sensitivity = specificity = 100%
  - In principle independent of the prevalence of the condition
  - **But they depend on the spectrum of patients (spectrum effect):** sensitivity is higher when cases are severe, specificity lower when non-cases have similar conditions
  - Report the setting and population in which they were estimated

Table: Cross-tabulation

		Gold standard +	Gold standard -
Test result	+	TP	FP
	-	FN	TN

TP = true positive
FP = false positive
FN = false negative
TN = true negative

18

AARHUS UNIVERSITY

Criterion validity, dichotomous instruments

Dichotomous gold standard

- Positive and negative predictive value (PPV/NPV)
  - PPV = TP/(TP+FP) = P(Sick | Positive test)
  - NPV = TN/(TN+FN) = P(Not sick | Negative test)
  - Population specific: PPV and NPV give information on how well a screening test will perform in a given population with known prevalence
  - Depend on the prevalence of disease in the population
  - Generally a higher prevalence will increase the PPV and decrease the NPV.

Table: Cross-tabulation

		Gold standard +	Gold standard -
Test result	+	TP	FP
	-	FN	TN

TP = true positive
FP = false positive
FN = false negative
TN = true negative

AARHUS UNIVERSITY

Criterion validity, dichotomous instruments

Dichotomous gold standard

- Example (Section 6.4.1)
  - Aim: To determine the diagnostic accuracy of MRI for the detection of breast cancer in the contralateral breast of a woman who had just been diagnosed with cancer in the other breast:
  - Instrument: MRI
  - Gold standard: Histological evidence of cancer within 1 year after the MRI.

Table: Cross-tabulation

		Gold standard +	Gold standard -
Test result	+	30	114
	-	3	822

AARHUS UNIVERSITY

Criterion validity, dichotomous instruments

Dichotomous gold standard

MRI vs. histology (gold standard), n = 969

	Histology +	Histology –	Total
MRI +	30	114	144
MRI –	3	822	825
Total	33	936	969

Conclusion

- MRI is very good at ruling out cancer: NPV 99.6%
- But about 4 of 5 positive MRIs are false positives: PPV 21%
- Wide CI for sensitivity – only 33 with cancer

Results

	Estimate	95% CI
Prevalence	3.4%	2.4–4.7%
Sensitivity	90.9%	75.7–98.1%
Specificity	87.8%	85.6–89.8%
PPV	20.8%	14.5–28.4%
NPV	99.6%	98.9–99.9%

Exact (Clopper–Pearson) CIs for proportions

AARHUS UNIVERSITY

Criterion validity, ordinal/continuous instruments

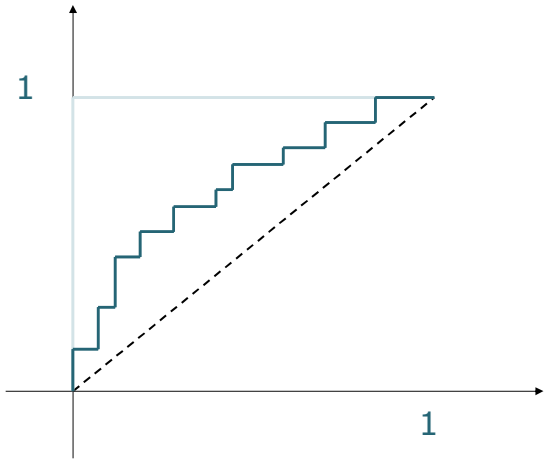
Dichotomous gold standard: the ROC curve

- A graph that plots sensitivity against 1-specificity. A dichotomous gold standard is held against all possible cut-offs on an ordinal or continuous measurement instrument.

AARHUS UNIVERSITY

Criterion validity, ordinal/continuous instruments

Dichotomous gold standard: area under the ROC curve (AUC)



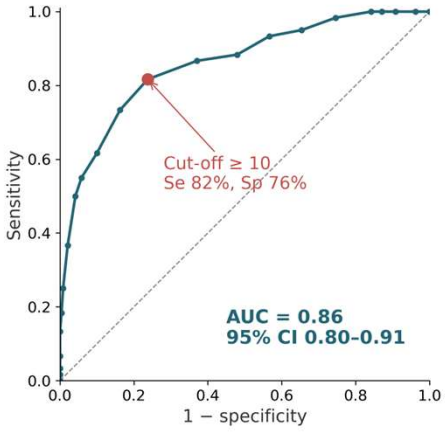
A measure of how well the measurement instrument performs against the gold standard – AUC
area under the curve
Range: 0–1 (0.5 = no better than chance)

23

AARHUS UNIVERSITY

Criterion validity, ordinal/continuous instruments

Dichotomous gold standard: ROC example



Questionnaire score (0–30) vs. clinical diagnosis
n = 300, of whom 60 have the condition (simulated data)

Cut-off	Sensitivity	Specificity	Youden J
≥ 8	88%	52%	0.40
≥ 9	87%	63%	0.50
≥ 10	**82%**	**76%**	**0.58**
≥ 11	73%	84%	0.57
≥ 12	62%	90%	0.52
≥ 13	55%	94%	0.49

- Youden index J = Se + Sp – 1 is highest at ≥ 10

24

AARHUS UNIVERSITY

Criterion validity, ordinal/continuous instruments

Dichotomous gold standard: choosing a cut-off and interpreting AUC

Choosing a cut-off

- **Youden index** (Se + Sp − 1): the point furthest from the diagonal
- **The purpose decides:**
  - to rule out the condition, few false negatives are needed → low cut-off, high sensitivity
  - to confirm the condition, few false positives are needed → high cut-off, high specificity
- Weigh the consequences of false positives against false negatives – and the prevalence
- A cut-off chosen in the same data is optimistic – confirm it in a new sample

Interpreting AUC

- The probability that a random person with the condition scores higher than a random person without
- 0.5 = chance, 1.0 = perfect
- **Rule of thumb:** 0.7–0.8 acceptable, 0.8–0.9 excellent, ≥ 0.9 outstanding
- **COSMIN:** AUC ≥ 0.70 for sufficient criterion validity
- Always report the 95% CI – in the example 0.86 (0.80–0.91), so the whole interval is above 0.70

Hosmer & Lemeshow, Applied Logistic Regression; Prinsen et al. 2018

AARHUS UNIVERSITY

Criterion validity, continuous instruments

Continuous gold standard

- Bland-Altman method
  - Here: the instrument compared with the gold standard (introductory course: the same instrument measured twice → measurement error)
  - Plot the difference against the average
  - Compare the two scores using a paired t-test
  - Calculate 95% limits of agreement ($\delta \pm 1.96 \cdot SD$)
  - Add horizontal lines at $y = 0$, $y = \delta$ (the mean difference) and at each of the limits of agreement
  - δ is a measure of the systematic difference (error) between the instrument and the criterion
  - $1.96 \cdot SD$ is a measure of the random error
  - Hence $\delta \pm 1.96 \cdot SD$ indicates the size of the measurement error

AARHUS UNIVERSITY

Criterion validity, continuous instruments

Continuous gold standard

- Bland-Altman plot and limits of agreement
  - Assumptions
    - The differences should be independent
    - The differences do not depend on the magnitude of the measurement (no trend in the plot)
    - The differences should be normally distributed

(see: Bland JM, Altman DG. *Measuring agreement in method comparison studies*. Stat Methods Med Res 1999;8(2):135–60)

27

AARHUS UNIVERSITY

Criterion validity, continuous instruments

Continuous gold standard

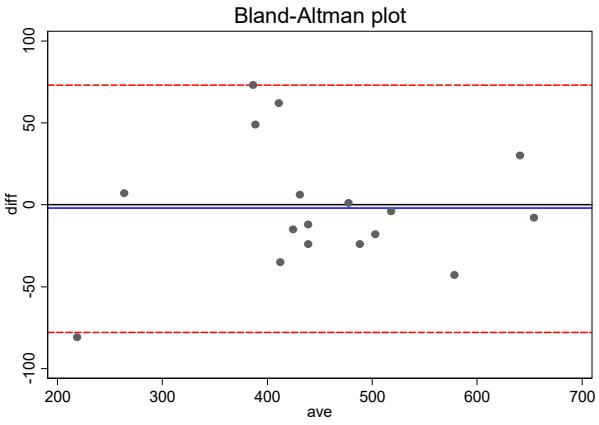
- Example
  - PEFR measurements on 17 persons
  - Each person measured twice; one time with Mini Wright peak-flow meter and one time using a Standard Wright peak-flow meter
  - Aim:
    - To assess criterion validity by comparing the test instrument (Mini Wright) to the criterion (Standard Wright)
    - If the two PEFR meters were unlikely to give readings which differed by less than, say, 10 l/min, we would have criterion validity

28

AARHUS UNIVERSITY

Criterion validity, continuous: example

Continuous gold standard



PEFR (l/min): Standard – Mini Wright, n = 17

	Result
Mean difference δ	−2.1
95% CI for δ	−22.0 to 17.8
SD of differences	38.8
Limits of agreement	−78.1 to 73.9
Paired t-test	p = 0.82

Conclusion

- No systematic difference between the meters
- But limits of agreement (±76 l/min) are far wider than the 10 l/min set in advance

→ **criterion validity not OK**

29

AARHUS UNIVERSITY

Criterion validity, predictive: example

CURB-65 in community-acquired pneumonia (Lim et al., Thorax 2003)

- **Instrument:** CURB-65 at admission – one point each for Confusion, Urea > 7 mmol/l, Respiratory rate ≥ 30/min, low Blood pressure, age ≥ 65 → score 0–5
- **Criterion – in the future:** death within 30 days
- **Data:** 1068 hospitalised patients, 30-day mortality 9%

CURB-65	30-day mortality	Management
0–1	1.5%	Home treatment
2	9.2%	Short admission
≥ 3	22%	Severe – hospital

30

AARHUS UNIVERSITY

Criterion validity, predictive: interpretation

Is it good?

- Clear gradient: mortality rises from 1.5% to 22% with the score
- AUC ≈ 0.78–0.82 in later cohorts – above the COSMIN criterion of 0.70
- **Good at ruling out:** few patients with a low score die → safe to treat at home
- **Poor at ruling in:** CURB-65 ≥ 3 has PPV ≈ 35% – most survive

→ **Sufficient for triage – not for predicting the individual patient's death**

Pitfalls in predictive validity

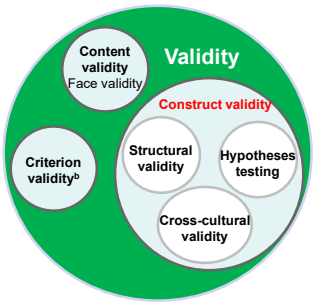
- **Treatment in between:** high-risk patients get more intensive treatment → lower observed mortality → predictive validity underestimated
- Loss to follow-up
- Validate in a new population before use

Lim et al., Thorax 2003; Barlow et al., Thorax 2007; Jones et al., Chest 2011

AARHUS UNIVERSITY

Construct validity

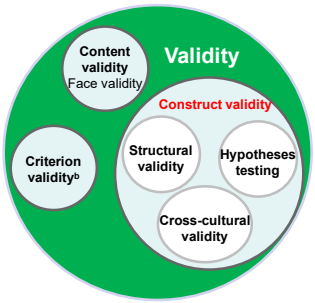
"The degree to which the scores of an instrument are consistent with hypotheses, e.g. with regard to internal relationships, relationships with scores of other instruments or differences between relevant groups"



AARHUS UNIVERSITY

Construct validity

- Used when there is no gold standard: we derive hypotheses about how the scores should behave if the instrument measures the construct – and test whether the data confirm them
- Three aspects
  - Structural validity
  - Hypotheses testing
  - Cross-cultural validity



AARHUS UNIVERSITY

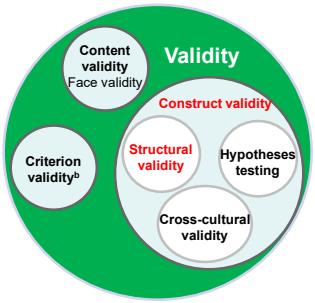
Structural validity

"The degree to which the scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured"

Assessed by

- **Confirmatory factor analysis (CFA):** preferred – tests a hypothesised structure
- **Exploratory factor analysis (EFA):** acceptable, but exploratory – mainly used during development
- **IRT/Rasch models:** unidimensionality, local independence, item fit

COSMIN Risk of Bias checklist (2018): CFA rated "very good", EFA "adequate"



AARHUS UNIVERSITY

Factor analysis

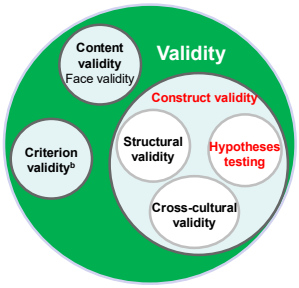
- Exploratory Factor Analysis (EFA)
  - Used when:
    - You have no idea of the number of factors
    - You have no idea of which items belong to each factor
- Confirmatory Factor Analysis (CFA)
  - Used when:
    - You know the number of factors (there should be in theory)
    - You know which items belong to each factor (theoretically)

More on EFA this afternoon: lecture and exercise

AARHUS UNIVERSITY

Hypothesis testing

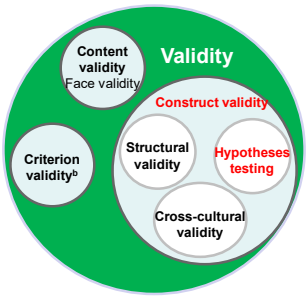
1. Describe the construct in high detail (use the conceptual model)
2. Formulate hypotheses about:
   - Positive correlations with similar constructs (Convergent validity)
   - No/slight correlation with unrelated constructs. (Discriminant validity).
   - Differences between subgroups of patients (Discriminative validity)
   - Focus on effect sizes and CI's; not statistical significance



AARHUS UNIVERSITY

Hypothesis testing

3. Describe the instruments to which your instrument is compared.
4. Describe characteristics of subgroups to be discriminated.
5. Gather empirical data
6. Test your hypotheses – are results consistent with hypotheses?
7. Discuss your findings



AARHUS UNIVERSITY

Hypotheses: direction and magnitude

Formulate hypotheses a priori – each with an expected direction and magnitude

Correlations with other instruments

convergent / discriminant validity

Comparator instrument measures …	Expected correlation
a similar construct	≥ 0.50
a related, but dissimilar construct	0.30 – 0.50
an unrelated construct	< 0.30

- Correlations with related constructs should be higher than with unrelated constructs – by at least 0.10
  - 0.10 is a convention: with n ≈ 100 a difference of this size is within sampling error – report it with a 95% CI

COSMIN user manual (Prinsen et al. 2018); de Vet et al. 2011; Cohen 1988

Differences between groups

discriminative validity

- Which group should score higher?
- By how much? E.g. as a standardised mean difference:
  - $d = (mean_1 - mean_2) / pooled\ SD$
  - 0.2 small, 0.5 moderate, 0.8 large (Cohen)

AARHUS UNIVERSITY

Hypotheses: evaluating the results

The 75% criterion

- **Sufficient construct validity:** ≥ 75% of the results are in accordance with the hypotheses
- **Insufficient:** < 75%
- Only meaningful if the hypotheses were formulated before the analysis

Effect sizes and 95% CI – not p-values

- **Effect size:** the size of the relationship or difference – r or ρ for correlations, d for group differences
- **95% CI:** the precision of the effect size – is the hypothesised magnitude supported by the whole interval, or only by the point estimate?
- p-values only test r = 0 or "no difference" – with a large n even trivial correlations are significant, and that is rarely the hypothesis

Terwee et al. 2007; COSMIN criteria for good measurement properties (Prinsen et al. 2018)

AARHUS UNIVERSITY

Correlations

Pearson r: strength of linear association

Spearman ρ: strength of monotone association (based on ranks)

Inference (test, CI) requires

- independent pairs (both)
- approx. bivariate normality (Pearson)

Watch out for

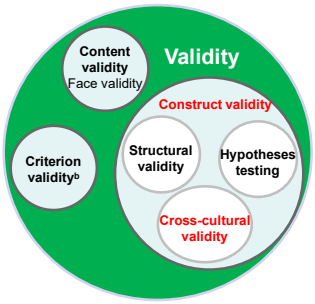
- non-linear relations: r ≈ 0 does not mean no relation
- outliers: one point can reverse the sign of r
- restricted range in the sample

Always plot the data!

AARHUS UNIVERSITY

Cross-cultural validity

"The degree to which the performance of the items on a translated or culturally adapted PRO instrument are an adequate reflection of the performance of items in the original version of the instrument".

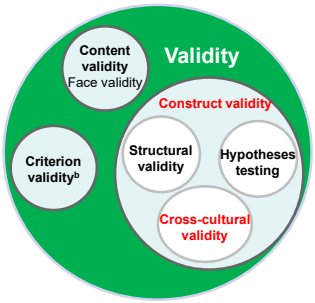


41

AARHUS UNIVERSITY

Cross-cultural validity

- Often assessed after translation of a questionnaire
- Checked by
  - Assessing its construct validity
  - Assessment of measurement invariance and differential item functioning (DIF) (CFA, logistic regression or IRT)
    - Measurement invariance: the items measure the construct in the same way in different groups (e.g. language versions), so scores can be compared
    - DIF: an item shows DIF if people from different groups with the same level of the construct answer it differently
    - More on DIF on day 3



A careful translation (e.g. Beaton et al. 2000) is a prerequisite – but cross-cultural validity must be tested with data: measurement invariance, DIF

42

AARHUS UNIVERSITY

Validation in context

Minimal n required

COSMIN Risk of Bias checklist (2018)

- Convergent/discriminant validity (correlations)
  - total n ≥100 (50–99 acceptable)
- Discriminative validity (t-test, rank test)
  - total n ≥100 (50–99 acceptable)
- Confirmatory factor analysis
  - ≥7 × number of items and ≥100
- Item response theory
  - Rasch/1PL models: ≥200
  - More complex models (e.g. 2PL): ≥1000

Missing values

- Report the percentage missing and how missing data were handled

AARHUS UNIVERSITY

Recap

- Essential that used measures/scores are valid – please spend the needed time to secure validity in your specific context!
- Many types of validity as illustrated in COSMIN circle.
- Depending on your context different types of validity will be possible to look into. If gold standard exists, criterion validity should be established.