Validity

Jens Søndergaard Jensen
M.Sc. in Statistics

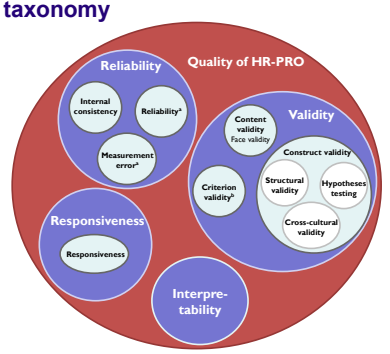
Outline

- Secrets to measuring a piece of paper
- The concept of validity
- Content and face validity
- Criterion validity
  - Concurrent validity
  - Predictive validity
- Construct validity
  - Structural validity
  - Hypotheses testing
  - Cross-cultural validity

Secrets to measuring a piece of paper

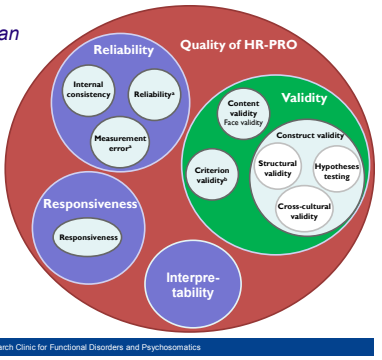
https://www.youtube.com/watch?v=9yUZTTLpDtk

The COSMIN taxonomy



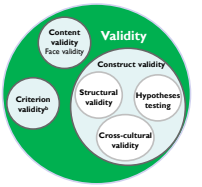
Validity

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



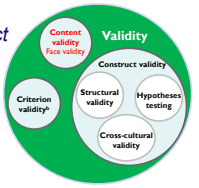
Validity

- Consists of testing hypotheses about the construct
- Important implications:
  - Knowledge about the construct
  - Complexity of the construct
  - Dependency on the situation
  - Validation of scores
    - Not measurement instruments
  - Formulation of specific hypotheses
  - Validation is a continuous process



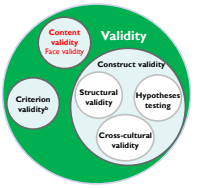
Content validity

- "*The degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured*"
  - Is the instrument relevant?
  - Is the instrument comprehensive?



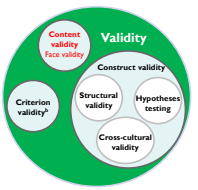
Face validity

- "*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



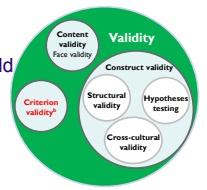
Content and face validity

- The proces 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 and comprehensive)
  5. Use a strategy or framework to assess correspondence between the instrument and construct



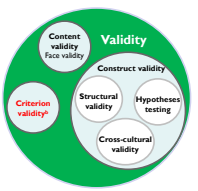
Criterion validity

- "*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?



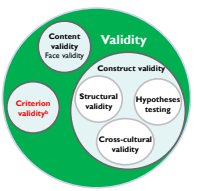
Criterion validity, subtypes

- 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 aplications



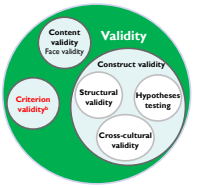
Criterion validity

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



Criterion validity

- The proces 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)

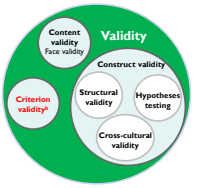


Criterion validity

Table 6.2: Overview of statistical parameters for various levels of measurement 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



Criterion validity, dichotomous instruments

- Sensitivity
  - Sens= TP/(TP+FN)
  - Probability of a postive 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	TP = true positive FP = false positive
	-	FN	TN	FN = false negative TN = true negative

Criterion validity , dichotomous instruments

- Sensitivity and specificity
  - A perfect test has Sensitivity = Specificity = 100%
  - They are characteristics of the test (and hence they are...)
  - Independent on the prevalence of disease in the population

Table: Cross-tabulation

		Gold standard +	Gold standard -	
Test result	+	TP	FP	TP = true positive FP = false positive
	-	FN	TN	FN = false negative TN = true negative

Criterion validity, dichotomous instruments

- Positive and negative predicted value (PPV/NPV)
  - PPV = TP/(TP+FP) = P(Sick | Positive test)
  - NPV = TN/(TN+FN) = P(Not sick | Negative test)
  - Both depend on the prevalence of disease in the population
  - Example: N=4000, Sens = Spec = 0.99
    - Prevalence = 50% ⇒ PPV = 99%
    - Prevalence = 10% ⇒ PPV = 92%
    - Prevalence = 5% ⇒ PPV = 84%
    - Prevalence = 1% ⇒ PPV = 52%

Table: Cross-tabulation

		Gold standard +	Gold standard -	
Test result	+	TP	FP	TP = true positive FP = false positive
	-	FN	TN	FN = false negative TN = true negative

Criterion validity, dichotomous instruments

- Example (Section 6.4.1)
  - Computations are easily done by hand
  - BUT: CI's are more difficult
  - Stata can do it for us
    - Download the `-diagt-` command

Table: Cross-tabulation

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

[redacted]

Criterion validity, dichotomous instruments

```
. diagti 30 3 114 822

       True |
    disease |     Test result
     status |    Neg.       Pos. |     Total
 -----------+----------------------+----------
     Normal |     822        114 |       936
   Abnormal |       3         30 |        33
 -----------+----------------------+----------
      Total |     825        144 |       969

                                                    [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                          Pr(A)      3.4%       2.4%      4.75%
---------------------------------------------------------------------------
Sensitivity                       Pr(+|A)     90.9%      75.7%     98.1%
Specificity                       Pr(-|N)     87.8%      85.6%     89.8%
ROC area                   (Sens. + Spec.)/2    .894       .843      .945
---------------------------------------------------------------------------
Likelihood ratio (+)      Pr(+|A)/Pr(+|N)      7.46       6.09      9.14
Likelihood ratio (-)      Pr(-|A)/Pr(-|N)      .104      .0352      .305
Odds ratio                      LR(+)/LR(-)     72.1         23       226
Positive predictive value         Pr(A|+)     20.8%      14.5%     28.4%
Negative predictive value         Pr(N|-)     99.6%      98.9%     99.9%
---------------------------------------------------------------------------
```

Table: Cross-tabulation

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

[redacted]

Criterion validity, continuous instruments

- Bland-Altman method
  - Plot the difference against the average
  - Compare the two scores using a paired t-test
  - Calculate 95% limits of agreement (δ ± 1.96·SD)
  - Add horizontal lines at y = 0, y = δ (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·SD is a measure of the random error
  - Hence δ ± 1.96·SD indicates the size of the measurement error

[redacted]

Criterion validity , continuous instruments

- Bland-Altman plot and limits of agreement
  - Assumptions
    - The differences should be independent
    - The differences should have the same distribution
    - The differences should be normally distributed
- One article you should read:
  - Bland JM, Altman DG. *Measuring agreement in method comparison studies*. Stat Methods Med Res 1999;8(2):135–60.

[redacted]

Criterion validity, continuous instruments

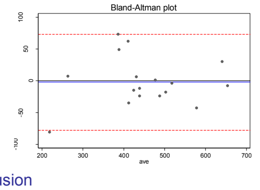
- Example
  - PEFR measurements on 17 persons
  - Each person measured twice; one time with Mini Wright peak-flow meter and one time using a Standard Wright peak-flow meter
  - Aim:
    - To asses criterion validity by comparing the test instrument (Mini Wright) to the criterion (Standard Wright)
  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

[redacted]

Criterion validity, continuous instruments

- Example

```
. use http://www.stata-press.com/data/mlmus3/pefr, clear

.
. * generate differences
. gen diff=wp1-wm1

.
. * generate averages
. gen ave=(wp1+wm1)/2

.
. * Limits of agreement (prediction interval)
. centile diff, centile(2.5 97.5) meansd

                                   -- Normal, based on mean and std. dev.--
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
        diff |      17         2.5     -78.09591       -110.2293   -45.96248
             |                97.5      73.86061        41.72719     105.994

.
. * paired t-test
. ttest wp1=wm1

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     wp1 |      17    450.3529    28.20995    116.3126    390.5505    510.1554
     wm1 |      17    452.4706    27.43443    113.1151    394.3122     510.629
---------+--------------------------------------------------------------------
    diff |      17   -2.117647    9.401925    38.76513   -22.04884    17.81354
------------------------------------------------------------------------------
     mean(diff) = mean(wp1 - wm1)                                 t =  -0.2252
 Ho: mean(diff) = 0                              degrees of freedom =       16

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.4123         Pr(|T| > |t|) = 0.8246          Pr(T > t) = 0.5877

.
. * Bland-Altman plot
. scatter diff ave, yline(0, lcol(black)) yline(-2, lcol(blue)) ///
>                   yline(-78 73, lcol(red) lpat(-)) scheme(s1mono) ///
>                                 name(bap, replace) title("Bland-Altman plot")
```

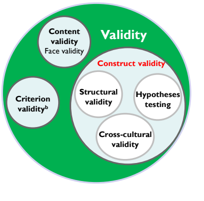


- Conclusion
  - δ = -2.1, 95% CI (-22.0 ; 17.8)
  - Limits of agreement (-78.1 ; 73.9)
  - Criterion validity not OK

[redacted]

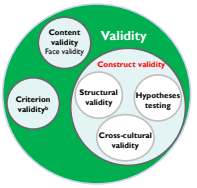
Construct validity

”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”



Construct validity

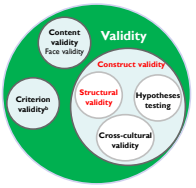
- Assumes the instrument validly measures the construct
- Three aspects
  - Structural validity
  - Hypotheses testing
  - Cross-cultural validity



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)
- Item Response Theory (IRT) (Indirectly)

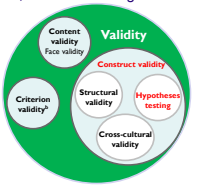


Factor analysis

- 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)
  - A statistical model
  - Use fit-parameters to assess whether data fit a hypothesized factor structure
  - Possible to compare fit of several models
  - Used when:
    - You know the number factors (there should be in theory)
    - You know which items belong to each factor (theoretically)

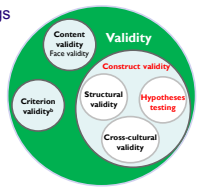
Hypothesis testing

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



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



Convergent/discriminative validity

- Example
  - Validation of several questionnaires used to assess functional health status in children with Acute Otitis Media (AOM)
  - 383 children with AOM
  - 7 questionaires assessing functional health status

Generic
- SF-36 (functioning and emotional behavior)
- FSQ generic (age appropriate functioning and emotional behavior)
- FSQ specific (general impact of illness on --------||--------)

Diseases specific
- OM-6 (physical functioning)
- NRS child (HRQoL of child)
- NRS caregiver (HRQoL of caregiver)
- FFQ (Family functioning)

Convergent/discriminant validity

- Hypotheses
  - Weak correlation (r = 0.1 - 0.3) between FSQ generic and NRS caregiver
  - Moderate to strong correlation (r > 0.4) between SF-36 and NRS caregiver
  - Moderate to strong correlation (r > 0.4) between OM-6, FSQ specific, NRS child, NRS caregiver and FFQ

Table 6.5: Construct validity: correlations between questionaires

	RAND	FSQ Generic	FSQ Specific	OM-6	NRS child	FFQ	NRS caregiver
RAND	1.00	**0.52**	**0.49**	**0.34**	**0.33**	**0.43**	**0.49**
FSQ Generic		1.00	**0.80**	**0.37**	0.25	**0.43**	**0.24**
FSQ Specific			1.00	**0.49**	0.26	**0.53**	0.24
OM-6				1.00	0.23	**0.74**	0.28
NRS child					1.00	0.22	**0.47**
FFQ						1.00	0.39
NRS caregiver							1.00

Convergent/discriminant validity

- Conclusions
  - NRS child does not perform as hypothesized
  - NRS caregiver shows lower correlation, with FSQ specific and OM-6, than hypothesized

Table 6.5: Construct validity: correlations between questionaires

	RAND	FSQ Generic	FSQ Specific	OM-6	NRS child	FFQ	NRS caregiver
RAND	1.00	**0.52**	**0.49**	**0.34**	**0.33**	**0.43**	**0.49**
FSQ Generic		1.00	**0.80**	**0.37**	0.25	**0.43**	**0.24**
FSQ Specific			1.00	**0.49**	0.26	**0.53**	0.24
OM-6				1.00	0.23	**0.74**	0.28
NRS child					1.00	0.22	**0.47**
FFQ						1.00	0.39
NRS caregiver							1.00

Discriminative validity

- Hypotheses:
  - Children with ≥4 episodes of AOM per year (n=242) have lower scores on all instruments than children with only two and three episodes per year (n=141)
    - NO magnitude was specified

Table 6.6: Discriminative validity: scores of children with 2-3 versus 4 or more episodes of AOM in the previous year

	2-3 AOM episodes	≥4 AOM episodes	p-value
Generic	21.1	19.6	0.004
RAND SF-36	76.5	72.2	0.002
FSQ Generic	83.9	78.4	0.001
FSQ Specific			
Disease specific			
OM-6	18.9	17.0	<0.001
NRS child	5.2	5.4	0.48
FFQ	84.9	78.5	<0.001
NRS caregiver	6.6	6.2	0.22

P-values calculated by Mann-Whitney test

Discriminative validity

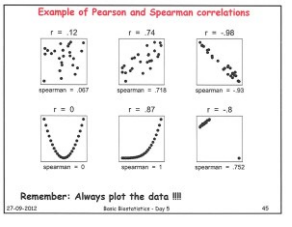
- Conclusions:
  - Statistical significant difference between the two groups on all instruments exept the two NRS questionaires
- Overall conclusions:
  - The NRS questionaires does not perform as well as expected
  - Results from convergent/discriminative validity support each other
  - Only the NRS questionaires showed poor covergent validity and low to moderate discriminative validity

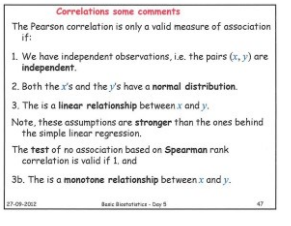
A warning about correlations

$$r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

- Pearson's r
  - Measures strength of a linear association between two continuous normally distributed variables
- Spearman's ρ
  - Measures strength of a monotone association between two variables
  - The non-parametric version of Pearson's r. Uses the rank of observations instead of the actual bservations
- Statistical significance (ie. ρ ≠ 0) is rarely important
- Remember warning from Basic Biostatistics

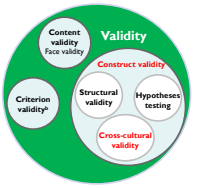
Correlations





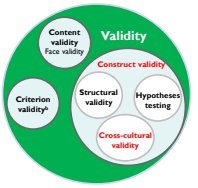
Cross-cultural validity

"*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*".



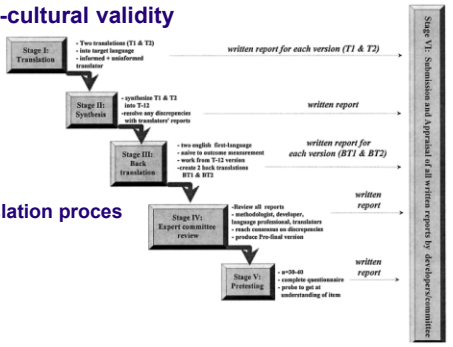
Cross-cultural validity

- Often assessed after translation of a questionaire
- Checked by
  - Assessing its construct validity
  - Assessment of measurement invariance, differential item functioning (DIF) (CFA, logistic regression or IRT)



Cros-cultural validity

The translation proces



(Beaton et al. *Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures.* Spine, Vol. 25, s. 3186-91)

Validation in context

- Minimal n required
  - Convergent/discrimant validity (correlations)
    - 50 patients (preferrebly over 100 patients)
  - Discriminative validity (t-test, rank test)
    - 50 patients per subgroup
  - Confirmatory Factor Analysis
    - 4-10 cases per item (at least 100 patients)
  - Item Response Theory
    - >200 to create stable models
    - Too many is also a problem
- Missing values
  - Should, as always, be reported and investigated
  - More than 15% missing → problems with generalizability