Measurement… again

Sound metrics

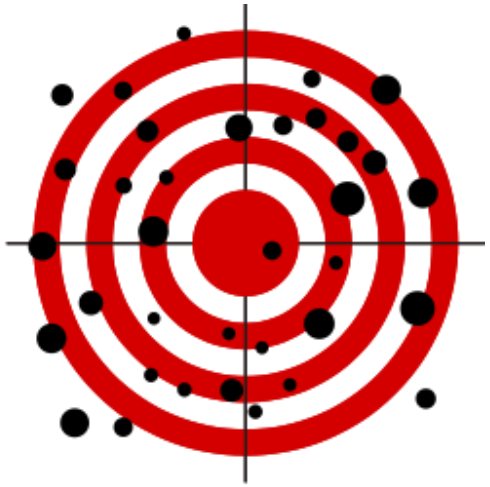
Reliability

- Degree of consistency
- Getting similar answers

Validity

- Degree of reflecting truth
- Getting correct answers

Unreliable & Unvalid



Unreliable, But Valid

Reliable, Not Valid

Both Reliable & Valid

We want tools that are consistent and do what they are supposed to do.

Reliabilities

- Inter-rater/inter-observer
  - Agreement between raters/observers
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation

	Rater 1	Rater 2	Agreement?
Item 1	Blue	Red	0
Item 2	Blue	Blue	1
Item 3	Red	Red	1
Item 4	Blue	Blue	1
Item 5	Red	Red	1
		Reliability	**4/5 = .80**

	Rater 1	Rater 2	
Item 1	1	3	
Item 2	3	5	
Item 3	2	2	
Item 4	5	7	
Item 5	2	1	
		Reliability	**r = .83**

Reliabilities

- Inter-rater/inter-observer
  - Agreement between raters/observers
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation
- Test-retest
  - Agreement between time points
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation

	Time1	Time 2	Agreement?
P 1	Blue	Red	0
P 2	Blue	Blue	1
P 3	Red	Red	1
P 4	Blue	Blue	1
P 5	Red	Red	1
		Reliability	**4/5 = .80**

	Time1	Time 2	
P 1	10	30	
P 2	30	50	
P 3	20	20	
P 4	50	70	
P 5	20	10	
		Reliability	**r = .84**

Reliabilities

- Inter-rater/inter-observer
  - Agreement between raters/observers
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation
- Test-retest
  - Agreement between time points
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation
- Parallel-forms
  - Agreement between equivalent variants
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation

	Form 1	Form 2	Agreement?
P 1	Blue	Red	0
P 2	Blue	Blue	1
P 3	Red	Red	1
P 4	Blue	Blue	1
P 5	Red	Red	1
		Reliability	**4/5 = .80**

	Form 1	Form 2	
P 1	10	30	
P 2	30	50	
P 3	20	20	
P 4	50	70	
P 5	20	10	
		Reliability	**r = .84**

Reliabilities

- Inter-rater/inter-observer
  - Agreement between raters/observers
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation
- Test-retest
  - Agreement between time points
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation
- Parallel-forms
  - Agreement between equivalent variants
  - Calculation
    - Categorical – Percentage in agreement
    - Continuous – Correlation

Internal Consistency

- Inter-item correlation
  - Average correlation of all items in the measure/test

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Item 5*	*Item 6*
Item 1	1					
Item 2	.89	1				
Item 3	.91	.92	1			
Item 4	.88	.93	.95	1		
Item 5	.84	.86	.92	.85	1	
Item 6	.88	.91	.95	.87	.85	1

Average correlation = $\frac{.89+.91+.88+.84+.88+.92+.86+.91+.95+.92+.95+.85+.87+.85}{15} = .89$

Internal Consistency

- Inter-item correlation
  - Average correlation of all items in the measure/test
- Item-total (part-whole) correlation
  - Average correlation of all items to total of measure/test

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Item 5*	*Item 6*
Item 1	1					
Item 2	.89	1				
Item 3	.91	.92	1			
Item 4	.88	.93	.95	1		
Item 5	.84	.86	.92	.85	1	
Item 6	.88	.91	.95	.87	.85	1
Part-Whole	**.84**	**.88**	**.86**	**.87**	**.83**	**.82**

Average correlation = $\frac{.84+.88+.86+.87+.83+.82}{6} = .85$

Internal Consistency

- Inter-item correlation
  - Average correlation of all items in the measure/test
- Item-total (part-whole) correlation
  - Average correlation of all items to total of measure/test
- Cronbach's alpha
  - Average correlation of all possible split samples.

Item 1	**Item 3**	**Item 4**
.87		
Item 2	Item 5	Item 6

Item 1	**Item 3**	**Item 4**
.87		
Item 2	Item 5	Item 6

Item 1	**Item 2**	**Item 3**
Item 4	Item 5	Item 6

Item 1	**Item 2**	**Item 4**
Item 3	Item 5	Item 6

Item 1	**Item 2**	**Item 5**
Item 3	Item 4	Item 6

Item 1	**Item 2**	**Item 6**
Item 3	Item 4	Item 5

Item 1	**Item 3**	**Item 5**
Item 2	Item 4	Item 6

Item 1	**Item 3**	**Item 6**
Item 2	Item 4	Item 5

Item 1	**Item 4**	**Item 5**
Item 2	Item 3	Item 6

Item 1	**Item 4**	**Item 6**
Item 2	Item 3	Item 5

Item 1	**Item 5**	**Item 6**
Item 2	Item 3	Item 4

Internal Consistency

- Inter-item correlation
  - Average correlation of all items in the measure/test
- Item-total (part-whole) correlation
  - Average correlation of all items to total of measure/test
- Cronbach's alpha
  - Average correlation of all possible split samples.

$$r_{kk} = \frac{k}{k-1}\left(1 - \frac{\sum {s_i}^2}{{s_t}^2}\right)$$

In which:

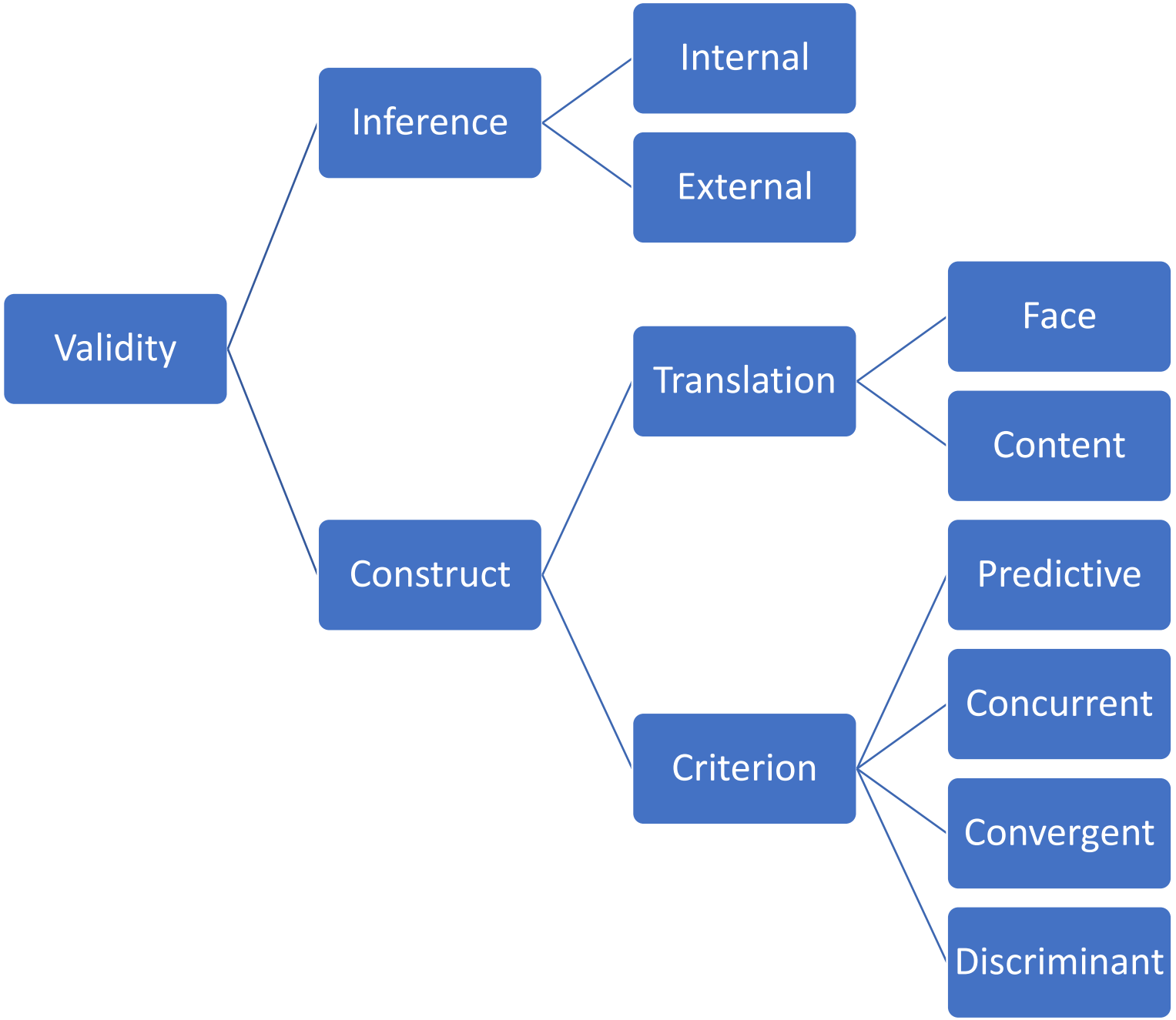
r_{kk}	=	the computed Cronbach's alpha
k	=	the number of items
${s_i}^2$	=	the variance of every item
${s_t}^2$	=	the variance of the total scale

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Number of items

Average variance

Average inter-item covariance among the items


Validity
Inference
Internal
External
Construct
Translation
Face
Content
Criterion
Predictive
Concurrent
Convergent
Discriminant

Inference validities

- Internal
  - Does my design allow me to draw the kinds of conclusions I did?
  - Do I have enough control?
- External
  - Do my results generalize?
  - Will the relationships I found hold in the real world?

Construct validities 1: Translation

- Face validity
  - Does it look like it is measuring the thing I am interested in, on the FACE of it?
- Content validity
  - Does it look like I adequately sampled the content necessary?

Construct validities 2: Criterion

- Predictive validity
  - Can it predict?
- Concurrent validity
  - Does it measure up to other measures of the same thing when administered concurrently?
- Convergent validity
  - Does it relate to stuff it should relate to?
- Discriminant validity
  - Does it **not** relate to stuff it shouldn't?

Calculating validities

- Translational
  - Subjective
- Criterion
  - Objective… ish
    - Use correlations, but what is enough? What is too much?

- Another example of more art than science.