

## EXS 511

### Reliability and Validity

Class 14 – May 8, 2011

Joy L. Hendrick, Ph.D.  
SUNY Cortland



## Validity

- “soundness of interpretation of a test” (Thomas, Nelson, & Silverman, 2005, p. 193)
- Does a test measure what its supposed to measure?
- Is the interpretation made from the test results reasonable? How true is it?
- Important when reviewing other studies
  - Why have they chosen specific tests/measurements?
- Important when you as the researcher select a test
- How do you know if it is valid?
  - Its a matter of degree.

## Types of Validity

- Logical/face validity
  - Does it appear to be a good measure?
  - Important starting point
- Content Validity
  - Does it include all the important elements?
- Construct Validity
  - Characteristic may not be directly observable
  - Do the sum of the constructs equal the whole?
    - Can examine differences in groups, or before and after a treatment

## Criterion Validity

- Is there a known valid (or truthful) measure? How does your test compare?
- Concurrent Validity
  - Relate/correlate to other known criterion(s) now
  - i.e. judges/expert ratings, other test
- Predictive validity
  - Will it predict a criterion at a future time?
  - Can the predictor scores predict the criterion?

## How do we measure/assess validity... statistically?

- Content validity
  - Are all the elements included?
    - First examine subjectively
    - Statistically can use factor analysis, cluster analysis linear modeling
    - i.e. Physical Fitness example, surveys, written tests
- Construct Validity
  - If all the parts equal the whole than one should find:
    - Differences between known groups (t-tests, ANOVA)
    - Differences before and after treatment
      - Dependent t-tests or repeated measures ANOVA

## Concurrent Validity

- Correlate scores with scores from other known criterion (also multiple regression)
- Pearson's  $r_{xy}$  can be validity coefficient
- Shoot for  $r_{xy} \geq .80$
- Samples include:
  - Other known valid test (i.e. underwater weighing for % body fat)
  - Expert ratings
- Could have problems finding a good (valid, reliable) criterion score

## Sit and Reach (SR) Study

- Lemmink, Kemper, de Greef, Rispen and Stevens (2003)
- Used American Academy of Orthopedic Surgeons procedures for ROM as **criteria**s of hamstring flexibility and lower back flexibility
- Looked at correlations
- Used stepwise multiple regression
- Compared shared variance explained by both regular and modified SR for older men and women.
- Found moderate support for SR to assess flexibility, but not for lower back flexibility



## Predictive Validity

- Correlate scores with scores from other known criterion at a future time (can also use multiple regression)
- Shoot for  $r$  or  $R \geq .80$
- i.e. future injury rates, future GPA, future test scores on national exams

## Other considerations

- If using multiple regression for validation, it tends to be population specific
- Cross-validation
  - Generate new sample from same parent population
  - Apply regression equation to verify if it is still significant

## Reliability

- Indication of consistency or stability of measurement
- Necessary ingredient for validity
  - A test must be reliable for it to be valid
  - If a test can not give you the same score time after time, how can it be a *valid* measure of that characteristic?

## Measurement Error

- Observed score = true score + error score

$$X = t + e$$

- Measurement error may be attributed to following factors:
  - Lack of agreement or consistency among scorers
  - Lack of consistent performance by the individual tested
  - Failure of instrument to measure consistently
  - Failure of tester to follow standardized testing procedures

## Measuring Reliability

- Degree of reliability expressed as a correlation (0 – +1.00)
- Interclass correlation (i.e. Pearson)
  - Problems in assessing reliability,  $r_{xx}$
- Intraclass reliability
  - Can be obtained from a repeated measures ANOVA (or from SPSS)

$$R = (MS_s - MS_e) / MS_s \text{ where}$$

$$MS_e = \frac{SS_T + SS_R}{df_T + df_R}$$

Therefore, do NOT use this for reliability!

## SPSS Example

- Test scores – given 3 times to a sample of participants (data file: reliability.sav)
- How reliable (stable) are the scores?
- Repeated measures ANOVA results
- Calculate  $MS_E$  and then R
- Effect of amount of score spread on reliability
- Baumgartner (2000) – good overview of stability

## Calculating R from repeated measures ANOVA

Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
trial	450	2	.225	.182	.834
Greenhouse-Geisser	450	1.263	.356	.182	.729
Huynh-Feldt	450	1.286	.350	.182	.734
Lower-bound	450	1.000	.450	.182	.672
Error(trial)	96.217	78	1.234		
Sphericity Assumed	96.217	78	1.234		
Greenhouse-Geisser	96.217	49.273	1.953		
Huynh-Feldt	96.217	50.161	1.918		
Lower-bound	96.217	39.000	2.467		

Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	151443.075	1	151443.076	2395.806	.000
Error	2465.258	39	63.212		

## SPSS Example

- Using Reliability Analysis results in SPSS to get R directly
  - Intraclass correlation R
- Analyze → scale → reliability analysis
  - Move over the 3 trials
  - Scale label = trials
  - Statistics – check
    - item statistics,
    - Intraclass correlation
      - Model = 2-way mixed

## Intraclass Correlation

Item Statistics

	Mean	Std. Deviation	N
trial1	35.53	4.723	40
trial2	35.60	4.673	40
trial3	35.45	4.641	40

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig.
Single Measures	.944 <sup>a</sup>	.907	.968	51.244	39.0	78	.000
Average Measures	.980 <sup>b</sup>	.967	.989	51.244	39.0	78	.000

a. Two-way mixed effects model where people effects are random and measures effects are fixed.  
b. Two-way mixed effects model where measures effects are random and people effects are fixed.

If you plan on using only one score,  $R = .944$

If you plan on using the mean of the 3 scores,  $R = .980$

## Measures of Establishing Reliability

- Stability
  - Stability of the test measure from day to day
  - Use test/retest method
  - Suggested 1-7 days as long as “ability” has not changed
- Alternate Forms (parallel forms) Method
  - Give both tests to a sample

- Internal Consistency
  - Same day test-retest
  - Split half or Kuder-Richardson (may see on some testing software)
  - Multiple trial tests and other scales– Cronbach Alpha (mean of all correlations)
    - SPSS – Alpha
    - Want .7 or higher

Acceptable reliability levels ? .70 is minimum

- if  $R > .80$ , 50-100 subjects is adequate
- if  $R < .80$ , should have > 100 subjects

(Baumgartner et al., 2003)

## SPSS example for questionnaire data

- File scalev3.sav
- Explained on p. \_\_\_\_\_
- Attitudes towards recycling
- Likert scale: 1 (strongly agree) to 5 (strongly disagree)
- Some items had reverse scoring – have been corrected (variables ending in R)
- Conduct alpha and split half reliabilities
- Analyze → Scale → Reliability Analysis

## Internal consistency results

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.899	.907	20

Cronbach's alpha for the ATR scale from the current sample was acceptable at .899

Item-Total Statistics

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
q_a	72.85	107.106	.546	.726	.894
q_c	72.92	106.418	.465	.642	.897
q_d	72.54	108.041	.605	.783	.893
q_e	72.35	108.617	.618	.700	.894
q_f	72.79	101.147	.710	.751	.889
q_j	72.40	105.148	.633	.610	.892
q_n	73.13	114.707	.062	.380	.910
q_o	72.58	108.248	.530	.821	.895
q_s	73.46	112.466	.240	.378	.902
q_t	72.94	108.860	.467	.785	.897
q_bR	72.85	106.766	.515	.720	.895

Item analysis on question N indicate that it may need to be replaced as it is inconsistent with rest of scale

## Intertester Reliability OBJECTIVITY

- Examines consistency of scoring across raters or judges
- Have more than one tester gather data
- If have multiple coders:
  - Inter-observer agreement (IOA)
 
$$IOA = \frac{\#agreements}{\#agreements + \#disagreements}$$
  - Cohen's Kappa
    - looks at level of agreement between 2 raters
    - And takes into account the variability in agreement
    - Excellent ( $\geq .75$ ); Fair to Good (.40-.74); Poor ( $< .40$ )

## Example

"Intertester Reliability and Validity of Motion Assessments During Lumbar Spine Accessory Motion Testing"\*

Kappa = .71 good

Table 1.

Intertester Reliability for Least Mobile Segment\*

Tester 2	Level	Tester 1					Total
		5	4	3	2	1	
5	5	0 (0)					0
4	4		1 (0.1)	1			2
3	3			2 (0)			2
2	2				7 (3.0)	3	10
1	1					14 (8.8)	15
Total		0	1	3	8	17	29

\* Level indicates the lumbar spinal vertebra to which the posterior-anterior force was applied. Values in parentheses indicate the calculated frequencies of agreements expected by chance. Agreement = 82.8%, kappa = .71, 95% confidence interval = .48 to .94.

\* Landel, Kulig, Fredericson, Li & Powers (2008) in *Physical Therapy*, 88 (1), 43-49

Table 2.

Intertester Reliability for Most Mobile Segment\*

Kappa = .29 poor

Tester 2	Level	Tester 1					Total
		5	4	3	2	1	
5	5	21 (19.9)	3				24
4	4	2	2 (0.7)				4
3	3			0 (0)			0
2	2	1			0 (0)		1
1	1					0 (0)	0
Total		24	5	0	0	0	29

\* Level indicates the lumbar spinal vertebra to which the posterior-anterior force was applied. Values in parentheses indicate the agreements expected by chance. Agreement = 79.3%, kappa = .29, 95% confidence interval = -.13 to .71.

Kappa can also be used for assessing validity

Kappa = .04 Very poor; Procedure is not valid at all

Table 3.

Validity for Least Mobile Segment\*

NIB	Level	Tester					Total
		5	4	3	2	1	
5	5	0 (0)			3	1	4
4	4		2 (0.6)		4	2	8
3	3			1 (0.3)	1	3	5
2	2				0 (2.0)	4	5
1	1				3	4 (3.25)	7
Total		0	2	2	11	14	29

\* Level indicates the lumbar spinal vertebra to which the posterior-anterior force was applied. Values in parentheses indicate the calculated frequencies of agreements expected by chance. Agreement = 24.1%, kappa = .04, 95% confidence interval = -.16 to .24.

## Obtaining Kappa in SPSS

- Kappa is one of statistics obtainable in Crosstabs
- Instructions can be found at:

<http://wpe.info/vault/wood07/Wood07.pdf>