Inter-Rater Reliability Roundtable Presentation

Winifred Nweke, Ph.D. Clayton State University
Tasha Perkins, Ph.D. Shorter University
Kristy Brown, Ph.D. Augusta University

Spring 2019 GADA Meeting
April 29, 2019
Macon, Georgia

Reliability Overview – Classical Testing Theory

- What is Reliability? What is its relationship to error?
- In Classical Test Theory: X =T + E
 Equation 1
 - where X = a test score, T = True score, E = Error
- $\sigma_{\mathsf{T}}^2 = \sigma_{\mathsf{X}}^2 \sigma_{\mathsf{F}}^2$ Equation 2
- Reliability = Proportion of Variability in X that is accounted for by the Variability in the true scores T.

$$r = (\sigma_X^2 - \sigma_E^2) / \sigma_X^2$$
 Equation 3

 Reliability describes the precision of the measurement process; or broadly consistency of measurement

Discussion

- What types of Reliability have you been/should you be reporting on the instruments you use?
- ■Mhh3;
- What challenges are you facing or have you faced?

Types of Reliability and Associated Sources of Error

- Assume the same multiple choice test was administered to a set of students once, and then after one month; what factors could induce variability in students scores?
 - Test-Retest; Sources of error:
 - Internal Consistency; Sources of error:
- Now assume the test was an essay test and administered once. What would be additional source of error? What about a performance assessment?
 - Inter-Rater; Sources of error:
- Thus, it is important to have a well defined rubric and trained raters

Methods of Establishing INTER-RATER RELIABILITY

- Percent Agreement (T = X-E)
- Cohen's Kappa (T = X-E)
- Pearson Correlation Coefficient (T=X-E)
- Intraclass Correlation Coefficient (ICC) Generalizability Test Theory:
 - T = (P+R) E
- $Arr r = (\sigma_X^2 \sigma_E^2) / \sigma_X^2$ Equation 3 from Slide 2 is generalized to:
- $r = ((\sigma_P^2 + \sigma_R^2) \sigma_E^2) / (\sigma_P^2 + \sigma_R^2)$ Equation 4

Percent Agreement

		Rater 2					
		Pass	Fail	Total			
Rater 1	Pass	<mark>55</mark>	10	<mark>65</mark>			
	Fail	20	<mark>15</mark>	35			
	Total	<mark>75</mark>	<mark>20</mark>	100			
	Percent Agreement Percent Agreement = $P_o = ((55 + 15)/100) \times 100 = 70\%$ Kappa (K) = $(P_o - P_c)/(1-P_c)$ where $P_c = [(65/100)\times(75/100)] + [(35/100)\times(20/100)] = 0.58$ Kappa (K) = $(.758)/(1.0 - 0.58) = 0.12/0.42 = 0.285$						

Percent Agreement - Rubric, 4 Criteria

Student B Lesson Plan		Rater 2					
		1	2	3	4	TOTAL	
Rater 1	1	0	0	0	0	0	
	2	0	1	1	1	3	
	3	0	2	<mark>5</mark>	0	7	
	4	0	0	0	0	0	
	TOTAL	0	3	6	1	10	
	Percent Agi	<u>reement</u>					
	Percent Agreement Po = ((0+1+5+0)/10)) X 100 = 60% Kappa (K) = $(P_o - P_c)/(1-P_c)$ where P_c = $(0/10)X(0/10) + (3/10)X(3/10) + (7/10)(6/10) + (0/10)(1/10) = 0.51 Kappa (K) = (.651)/(1.0-0.51) = .09/0.49 = 0.18$						

Pearson Correlation Example

Students	Rater1	Rater 2	Rater 2-
			Rater1 scores
Jane	7	10	3
John	6	9	3
Jo	5	8	3
Jones	3	7	4
Mary	2	5	3
Percent Agreemen	t is 0%,		

Pearson r = 0.978; focusing on their rank ordering; ignores the systematic differences in ratings. Can you think when these systematic difference would not be a problem? Awards to the best; both raters would pick Jane.

Calculating Inter-Rater Reliability Using Intraclass Correlation (Arrange data as shown below)

	Variables (Raters)							
Indicators	Rater1	Rater2	Rater3		Rater _n			
Indicator1								
Indicator2								
•••								
Indicator 24								

CAEP EVALUATION FRAMEWORK FOR EPP-CREATED ASSESSMENTS, Jan 2017

	XAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAFD STIFFICIENT LEVEL		XAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
		 CAEP SUFFICIENT LEVEL 4. DATA RELIABILITY a. A description or plan is provided that details the type of reliability that is being investigated or has been established (e.g., test-retest, parallel forms, inter-rater, internal. consistency, etc.) and the steps the EPP took to ensure the reliability of the data from the assessment. b. Training of scorers and checking on inter-rater agreement and reliability are documented. c. The described steps meet accepted research standards for establishing reliability. 		_
d.	evidence, is provided that scorers are trained, and their inter-rater agreement is documented. Described steps do not meet accepted research standards for reliability.			reported.

Augusta University Example

- Dispositions-Intraclass correlation coefficient
 - 7 Raters
 - ■32 Indicators

Dispositions

- Norming Training
 - Scenario
 - Small group discussion
 - Whole group discussion
 - Review of specific indicators
 - Evaluation of the final dispositions using multiple scenarios (Chalk and Wire)

Indicator	Supervisor 1	Supervisor 2	Supervisor 3	Supervisor 4	Supervisor 5	Supervisor 6	Supervisor 7
1	1	1	0	1	1	1	2
2	1	0	1	1	1	2	1
3	1	1	1	1	1	1	1
4	2	2	2	2	2	2	2
5	0	0	0	0	1	0	0
6	1	0	1	0	1	0	1
7	2	2	0	2	1	0	2
8	3	3	3	3	3	3	3
9	0	0	0	0	0	0	0
10	2	2	3	2	3	2	2
11	1	1	1	1	1	1	1
12	3	2	0	2	2	2	2
13	0	0	0	0	0	0	0
14	1	1	1	0	0	0	1
15	1	2	2	1	1	1	1
16	0	1	0	1	1	1	1
17	3	3	2	3	3	3	3
18	2	0	0	3	3	0	2
19	3	3	3	3	3	3	3
20	2	3	3	3	3	2	3
21	2	3	2	3	3	3	3
22	2	3	3	2	3	3	3
23	3	3	0	3	3	3	3
24	1	3	2	2	3	0	3
25	0	1	0	0	0	0	1
26	2	2	0	2	2	2	3
27	1	1	2	0	1	0	1
28	0	0	0	0	0	0	1
29	1	1	0	0	0	0	1
30	1	3	0	2	3	3	2
31	2	3	3	2	3	3	2
32	2	3	2	2	3	1	2

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	32	88.9
	Excluded ^a	4	11.1
	Total	36	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.943	.946	7

Item Statistics

	Mean	Std. Deviation	N
Supervisor1	1.4375	.98169	32
Supervisor2	1.6563	1.18074	32
Supervisor3	1.1250	1.21150	32
Supervisor4	1.4688	1.13548	32
Supervisor5	1.7188	1.19770	32
Supervisor6	1.3125	1.22967	32
Supervisor7	1.7813	.94132	32

Inter-Item Correlation Matrix

	Supervisor1	Supervisor2	Supervisor3	Supervisor4	Supervisor5	Supervisor6	Supervisor7
Supervisor1	1.000	.718	.495	.852	.766	.711	.805
Supervisor2	.718	1.000	.617	.774	.796	.743	.830
Supervisor3	.495	.617	1.000	.472	.603	.514	.506
Supervisor4	.852	.774	.472	1.000	.907	.747	.884
Supervisor5	.766	.796	.603	.907	1.000	.741	.859
Supervisor6	.711	.743	.514	.747	.741	1.000	.702
Supervisor7	.805	.830	.506	.884	.859	.702	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1.500	1.125	1.781	.656	1.583	.055	7

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Supervisor1	9.0625	35.996	.828	.762	.934
Supervisor2	8.8438	33.555	.860	.772	.930
Supervisor3	9.3750	36.565	.592	.493	.955
Supervisor4	9.0313	33.709	.888	.905	.927
Supervisor5	8.7813	32.886	.902	.876	.926
Supervisor6	9.1875	33.899	.789	.646	.937
Supervisor7	8.7188	35.886	.881	.846	.930

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
10.5000	46.710	6.83445	7

Intraclass Correlation Coefficient

	Intraclass h			F Test with True Value 0				
	Correlation ^D	Lower Bound	Upper Bound	Value	df1	df2	Sig	
Single Measures	.703 ^a	.582	.816	17.600	31	186	.000	
Average Measures	.943 ^c	.907	.969	17.600	31	186	.000	

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

- a. The estimator is the same, whether the interaction effect is present or not.
- b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.
- c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

- Analyze
- Scale
- Reliability Analysis
- Bring all supervisors over
- Statistics
 - Descriptives Inter-item (select correlations)
 - Summaries (select means)
 - Intraclass correlation coefficient
 - Model = Two-Way Random because you are dealing with a sample of raters and each rater assessed all dimensions (or all 10 areas)
 - Select One-Way Random if you have large data sets and not all of your raters rated all areas (so if you had to divide the work up and rater 1 did not rate data 400-500).
 - Select Two-way Mixed if you have an entire population of raters/supervisors. Type = Consistency because you are interested in the mean rater scores (not individual rater/supervisor reliability)
 - Select Absolute Agreement if you're wanting to assess the reliability for a single individual. This will tell you how well that person's/rater's score assess the construct.
 - Okay
 - Run/Ok

Clayton State University Example

- Using Intra-class Correlation (ICC) to establish inter-rater reliability
- From Generalizability Theory Most flexible reliability index because it could be calculated in different ways to account for or ignore systematic score differences and works with ordered score categories too.
- ICC addresses the limitations of both agreement indices and Pearson correlations by separating the observed score X into more pieces than was done with the Classical theory. The pieces depend on how data were collected.

Steps for Calculating ICC Using SPSS

- Analyze>Scale>Reliability Analysis
- Select "Statistics Correlation coefficient"
- Check "intraclass"
 - Descriptive item, scale
 - Inter-item correlations
 - Summary

Intra-class correlation (ICC)

Model

- One-way Random (if raters are different for rates); each rater rates two
 candidates; counterbalanced so random two raters rate each candidate
- Two-way mixed (assumes raters are fixed i.e. using the population & subjects are random)
- Two-way random (models both rater and rate effects & both are drawn from large populations)

Type

- Absolute (absolute agreement on ratings, rather than consistency)
- Continue
- OK
- Interpret Average Measures under the Intra-class correlation coefficient

CSU Example

From A Norming Exercise, January 2018

12 Raters rating on 18 Indicators

English Lesson Video

Lesson Plan

Dispositions assessment

Debriefing notes etc.

Form 2 - Indicator By Rater Summary

VideoEngl	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6 F	Rater7	Rater8	Rater9	Rater10	Rater11	Rater12
Indctr01_E	2	2	2	1	2	3	3	0	2	2	1	2
Indctr02_E	1	2	2	3	2	1	2	2				1
Indctr03_E	1	2	3	2	2	3	2	2	2	2	3	2
Indctr04_E	2	2	0	2	2	0	2	2	0	2		
Indctr05_E	0	2	2	2	0	1	0	3	2	2	1	1
Indctr06_E	0	0	0	1	0	0	0	0	0	0	O	2
Indctr07_E	3	3	2	2	3	2	3	3	2	3	1	3
Indctr08_E	0	0	0	0	0	0	0	2	0	0	0	0
Indctr09_E	2	2	3	2	2	1	2	2	2	2	1	2
Indctr10_E	2	2	2	2	2	1	1	2	2	2	1	1
Indctr11_E	2	2	3	2	1	1	1	2	2	2	1	2
Indctr12_E	1	1	1	2	1	1	1	2	2	2	1	2
Indctr13_E	0	0	0	0	0	0	0	2	0	0	0	0
Indctr14_E	0	0	0	0	0	0	0	0	0	0	0	0
Indctr15_E	0	0	2	1	1	1	1	2	2	0	3	0
Indctr16_E	1	2	2	2	1	2	1	2	2	2	1	1
Indctr17_E	1	1	3	1	1	1	1	0	2	2	1	1
Indctr18_E	1	1	2	1	1	1	1	2	2	0	1	1

Item Statistics for English

	Item Statistics for English							
	Mean	Std. Deviation	N					
Rater1	1.06	.938	18					
Rater2	1.33	.970	18					
Rater3	1.61	1.145	18					
Rater4	1.44	.856	18					
Rater5	1.17	.924	18					
Rater6	1.06	.938	18					
Rater7	1.17	.985	18					
Rater8	1.67	.970	18					
Rater9	1.44	.922	18					
Rater10	1.39	1.037	18					
Rater11	1.00	.840	18					
Rater12	1.17	.924	18					

Inter-Item Correlation Matrix for English

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6	Rater7	Rater8	Rater9	Rater10	Rater11	Rater12
Rater1	1.000	.819	.460	.554	.872	.465	.817	.216	.446	.763	.149	.600
Rater2	.819	1.000	.600	.803	.788	.625	.739	.438	.614	.916	.289	.591
Rater3	.460	.600	1.000	.547	.510	.679	.478	.194	.898	.581	.612	.510
Rater4	.554	.803	.547	1.000	.645	.407	.535	.472	.630	.788	.409	.496
Rater5	.872	.788	.510	.645	1.000	.600	.937	.263	.530	.727	.455	.517
Rater6	.465	.625	.679	.407	.600	1.000	.690	.086	.718	.582	.597	.600
Rater7	.817	.739	.478	.535	.937	.690	1.000	.123	.497	.681	.426	.550
Rater8	.216	.438	.194	.472	.263	.086	.123	1.000	.307	.312	.289	.066
Rater9	.446	.614	.898	.630	.530	.718	.497	.307	1.000	.609	.608	.530
Rater10	.763	.916	.581	.788	.727	.582	.681	.312	.609	1.000	.270	.604
Rater11	.149	.289	.612	.409	.455	.597	.426	.289	.608	.270	1.000	.152
Rater12	.600	.591	.510	.496	.517	.600	.550	.066	.530	.604	.152	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1.292	1.000	1.667	.667	1.667	.049	12
Inter-Item Correlations	.535	.066	.937	.872	14.279	.043	12

Intraclass Correlation Coefficient for English

		95% Cor Inte		F	Test with T	True Value (0
	Intraclass	Lower	Upper				
	Correlation ^b	Bound	Bound	Value	df1	df2	Sig
Single Measures	.519 ^a	.354	.721	14.757	17	187	.000
Average Measures	.928 ^c	.868	.969	14.757	17	187	.000

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

- a. The estimator is the same, whether the interaction effect is present or not.
- b. Type A intraclass correlation coefficients using an absolute agreement definition.
- c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Shorter University Example – Disposition Assessment (CDPA)

- Norming Workshop:
 - Faculty were asked to "self-assess" with the rubric
 - Faculty scored 1 scenario and submitted scores
 - Each indicator was discussed and the group came to consensus
 - Any indicator with discrepant scores (not adjacent or exact) from original scores was considered in a following scenario
 - 4 scenarios completed
- Reliability Study: Tested Cronbach's alpha for each scenario and all indicators; Tested ICC [via website]
 - When using Cronbach's alpha for interrater reliability, the "variable" is the rater (MacLennon, 1993)

Covariance Matrix

22.9714	15.6286	13.9143	8.6	11.8	13.6	4.9143
15.6286	22.1714	16.8857	7.8	12.4	11.8	5.8857
13.9143	16.8857	20.7429	6.8	11.4	12.8	7.7429
8.6	7.8	6.8	10.4	6.2	9.4	5.8
11.8	12.4	11.4	6.2	13.6	8.2	6.4
13.6	11.8	12.8	9.4	8.2	20.4	8.8
4.9143	5.8857	7.7429	5.8	6.4	8.8	12.7429

Unstandardized Alpha

Alpha = 0.8992 n=35 SE=0.025 95%CI=0.8502 to 0.9482

Correlation Matrix

1	0.6925	0.6374	0.5564	0.6676	0.6282	0.2872
0.6925	1	0.7874	0.5137	0.7141	0.5548	0.3502
0.6374	0.7874	1	0.463	0.6787	0.6222	0.4762
0.5564	0.5137	0.463	1	0.5213	0.6454	0.5038
0.6676	0.7141	0.6787	0.5213	1	0.4923	0.4862
0.6282	0.5548	0.6222	0.6454	0.4923	1	0.5458
0.2872	0.3502	0.4762	0.5038	0.4862	0.5458	1

Standardized Alpha

Alpha = 0.9002 n=35 SE=0.0261 95%CI=0.849 to 0.9515

INTERRATER RELIABILITY STUDY, 2018-19, CDPA, 7 RATERS

Intraclass Correlations

ICC	Single	Meaned
Model 1	0.5428	0.8926
Model 2	0.5446	0.8933
Model 3	0.5602	0.8992

Questions?

Thank You

Contact Information:

Winifrednweke@clayton.edu

tperkins@shorter.edu

KBROWN15@augusta.edu

References & Resources

- Bajpai, S., Bajpai, R. & Chaturvedi, H. (2015). Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods. Journal of the Indian Academy of Applied Psychology, 41, 20-27. Retrieved from https://www.researchgate.net/publication/273451591 Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data An Overview of Concepts and Methods
- Bresciani, J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research, and Evaluation, 14*(12), 1-7.
- Knapp, T. R. (1991). Focus on Psychometrics. Coefficient alpha: Conceptualizations and anomalies. Research in Nursing & Health, 14(6), 457-480.
- MacLennan, R. N. (1993). Interrater Reliability with SPSS for Windows 5.0. The American Statistician, 47(4), 292-296.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved from https://pareonline.net/getvn.asp?v=9&n=4

Free Statistical Software/ Programs

- Website: http://www.obg.cuhk.edu.hk/ResearchSupport/StatTools/ResourceIndex Subjects.php
- PSPP to download: https://www.gnu.org/software/pspp/get.html