02a: Test-Retest and Parallel Forms Reliability

Quantitative Variables

1. **Classic Test Theory (CTT)**
2. **Correlation for Test-retest (or Parallel Forms): Stability and Equivalence for Quantitative Measures**
3. **Consistency vs Agreement**
4. **Intraclass Correlation Coefficient, ICC**
5. **ICC with SPSS**
6. **ICC Real Data**
7. **Comparison of Results with Menon**
8. **ICC with More than Two Assessment Periods or Forms**
9. **Single Item Test-retest**
10. **Published Examples of Test-retest (to be updated)**

Qualitative Variables

11. **Percent Agreement (see presentation "07a Coder Agreement for Nominal Data")**
12. **Nominal Variables: ICC, Kappa, Scott's Pi, Krippendorff's alpha (see "07a Coder Agree. for Nominal Data")**
13. **Ordinal Variables: Weighted Kappa (see "07b Coder Agreement for Ranked Data")**

Quantitative Variables

1. Classic Test Theory (CTT)

CTT tells us that when we attempt to measure something, like test anxiety, we understand that the score we observe, the observed score X, is made of two parts, a true score (T) and error (E):

$$X = T + E$$

We would like to know how much error, E, is included when we use observed scores, X, because the more error, the worse our measurement and the less confidence we have that X measures what we hope it measures.

Since there will almost always be variability in scores, we can say that the variance for scores will be greater than 0.00. If we use the symbol X for test anxiety scores, we can indicate the variance like this:

$$VAR(X)$$

We can also expect variance in both true scores, T, and error in measurement, E, so we can symbolize these variances too:

$$VAR(T) \text{ and } VAR(E)$$

Reliability is defined as the ratio of true score variance to observed score variance:

$$\text{Reliability, } rxx = \frac{VAR(T)}{VAR(X)}$$

Since X = T + E, we can show that reliability is the ratio of true score variance to true score variance plus error variance:

$$\text{Reliability, } rxx = \frac{VAR(T)}{VAR(T) + VAR(E)}$$

Reliability is the

- proportion of true score variance to observed score variance;
- should not be less than 0.00;
- should not be greater than 1.00;
- r or rxx or r_{xx} is the sample symbol for reliability,
- ρ or ρxx or ρ_{xx} is the population symbol for reliability, and
- unfortunately, both r and ρ are also symbols for Pearson correlation, so easy to confuse the two.

If there were no error in measurement, then VAR(E) would be zero, VAR(E) = 0.00, and reliability would be equal to 1.00:

$$= \frac{\mathrm{VAR(T)}}{\mathrm{VAR(T)} + \mathrm{VAR(E)}}$$

$$= \frac{\mathrm{VAR(T)}}{\mathrm{VAR(T)} + 0}$$

$$= \frac{\mathrm{VAR(T)}}{\mathrm{VAR(T)}} = 1.00$$

A reliability of 1.00 means no measurement error and therefore we have true scores.

Assumptions of CTT:

- Expected value of E = 0.00 (i.e., mean of errors will be 0.00)
- Covariance T and E = 0.00; Cov(T,E) = 0.00 (i.e., correlation of T with E = 0.00)
- Covariance Ej and Ek = 0.00, Cov(Ej,Ek) = 0.00 (i.e., correlation of Ej with Ek = 0.00)

In words, CTT indicates that measurement error, E, is random and therefore correlates with nothing; if E does show a correlation with something, it will likely be a weak correlation that is random (i.e., varies across samples and due to sampling variation).

Technical note:

VAR(X) = VAR(T) + VAR(E) + 2Cov(T,E)

Since E does not correlate with anything,

VAR(X) = VAR(T) + VAR(E) ~~+ 2Cov(T,E)~~

2. Correlation for Test-retest or Parallel Forms: Stability and Equivalence for Quantitative Measures

As a reminder, recall that **test-retest reliability** refers to situations in which an instrument is administered to participants, time elapses, then the instrument is re-administered to the same participants. Scores from both time periods are assessed to determine stability of scores. For **parallel forms reliability**, one administers two forms of an instrument, both designed to measure the same thing and provide the same scores for a given individual, to participants and then assess equivalence of scores. Both test-retest and parallel forms reliability follow the same mechanics and use the same reliability estimates, so the logic and estimation methods presented below apply equally to both test-retest and parallel forms.

According to CTT, the Pearson product moment correlation, r, is a measure of reliability between two parallel measures, or test-retest measures that provide quantitative scores:

$$\text{Pearson, r = Reliability, rxx} = \frac{\mathrm{VAR(T)}}{\mathrm{VAR(X)}}$$

In a testing situation for which test-retest or parallel forms reliability applies, if scores are measured without error, then one should obtain the same score for the same person on both administrations of the same instrument or parallel forms of the instrument.

The reliability coefficient of scores from Time 1 to Time 2 is known as coefficient of stability, or coefficient of equivalence if dealing with parallel forms. The means for scores from both Time 1 and 2 should be the same for perfect stability and equivalence. To the degree means differ, stability and equivalence is degraded so the measure of reliability should also diminish.

(Note: address weaknesses of test-retest designs)

The example below illustrates what should happen in test-retest if measurement occurs without error, and if scores do not change due to maturation, learning, or other changes to attitudes, conditions, etc.

Example 1: True Scores, Test Retest

Student	Test True Score	Re-test True Score
1	95	95
2	90	90
3	85	85
4	80	80
5	75	75
6	70	70
7	65	65
8	60	60

In the above example, true scores = observed scores, so

VAR(T) = VAR(X) = 140.00

Note, the variance above represents the total variance for both administrations of the test and retest, so 16 observations, not 8.

Reliability of these scores is

$$rxx = \frac{\mathrm{VAR(T)}}{\mathrm{VAR(X)}} = \frac{140.00}{140.00} = 1.00$$

The Pearson correlation for these two sets of scores is

r = 1.00

which indicates, that for these data, Pearson r validly estimates reliability for test-retest or parallel forms.

Example 2: True Scores with Error Added

Student	True Score	Error Time 1	Error Time 2	Observed Time 1 (True + Error 1)	Observed Time 2 (True + Error 2)
1	95	3	-3	98	92
2	90	-3	-3	87	87
3	85	3	3	88	88
4	80	-3	3	77	83
5	75	-3	3	72	78
6	70	3	3	73	73
7	65	-3	-3	62	62
8	60	3	-3	63	57

Note: Cov(e1,e2) = 0.00, Cov(e1,T) = 0.00, Cov(e2, T) = 0.00; errors uncorrelated with each other and true scores.

How well does Pearson r work if "random" measurement error is introduced to true scores?

Variances for true scores and observed scores in Example 2 are reported below.

VAR(T) = 140.00 (Variance of 16 true scores to mimic test and retest situation)
VAR(X) = 149.60 (Variance of both Time 1 and Time 2 observed scores combined)

The CTT reliability is

$$rxx = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{149.60} = 0.935$$

which means that 93.5% of variance in observed scores is due to true score variance, or 100(1 - .935) = 6.5% is error variance.

Pearson correlation for these data is

r = 0.935

(Note: Demonstrate for class that Pearson r obtained in SPSS or Excel is .935).

Results show that Pearson r works well to measure reliability when only random measurement error is included, and the means for both sets of scores are the same or similar. In Example 2 above, the means for Observed Time 1 = 77.50 and for Observed Time 2 = 77.50. However, Pearson r can fail when non-random error is included that changes means between the two sets of scores.

3. Consistency vs. Agreement

Consistency refers to the relative position of scores across two sets of scores. Consistency is an assessment of whether two sets of scores tend to rank order something in similar positions. **Agreement** refers to the degree to which two sets of scores agree or show little difference in actual scores; the lower the absolute difference, the greater the agreement between scores.

Pearson r is designed to provide a measure of consistency. Loosely described, this means Pearson r helps assess whether relative rank appears to be replicated from one set of scores to another.

Pearson r does not assess magnitude of absolute differences and can therefore present a misleading assessment of reliability when test-retest scores or parallel scores show large differences.

As Example 3 demonstrates, Pearson r shows a value of .91 for the Relative Reliability scores, but note that the actual scores are very different (Mean for Test 1 = 77.50, mean for Test 2 = 16.62).

Example 3: Relative vs. Absolute Reliability

	Relative Reliability, Consistency				Absolute Reliability, Agreement		
Student	Test 1	Rank 1	Test 2	Rank 2	Test 1	Test 2	Difference
1	95	1	44	1	95	92	3
2	90	2	22	2	90	91	-1
3	85	3	20	3	85	83	2
4	80	4	19	4	80	79	1
5	75	5	10	5	75	78	-3
6	70	6	9	6	70	72	-2
7	65	7	8	7	65	64	1
8	60	8	1	8	60	61	-1
	Test 1 and 2 Pearson r = .91				Test 1 and 2 Pearson r = .98		

Example 4 helps to solidify the problem with using Pearson r to assess test-retest and parallel forms reliability.

In Example 4, note that time 2 scores have error, but also has a growth component of 20 points from time 1. The two sets of observed scores, Time 1 and Time 2, are no longer equivalent, so scores are no longer stable over time.

Example 4: True Scores with Error and Systematic Difference Added

Student	True Score	Error Time 1	Error Time 2	Time 2 Change	Observed Time 1 (True + Error 1)	Observed Time 2 (True + Error 2 + Change)
1	95	3	-3	20	98	112
2	90	-3	-3	20	87	107
3	85	3	3	20	88	108
4	80	-3	3	20	77	103
5	75	-3	3	20	72	98
6	70	3	3	20	73	93
7	65	-3	-3	20	62	82
8	60	3	-3	20	63	77

Variances for true scores and observed scores:

VAR(T) = 140.00 (Variance of 16 true scores to mimic test and retest situation)
VAR(X) = 256.26 (Variance of both Time 1 and Time 2 observed scores combined)

The CTT reliability is

$$r_{xx} = \frac{\text{VAR(T)}}{\text{VAR(X)}} = \frac{140.00}{256.26} = 0.546$$

which means that 54.6% of variance in observed scores is due to true score variance.

The Pearson correlation, however, between Observed scores at Time 1 and 2, is

r = 0.935

(Note: Demonstrate for class that Pearson r obtained in SPSS or Excel is .935).

The Pearson r of .935 suggests the scores are stable over time and therefore provides a misleading assessment of stability.

In some situations, one desires a measure of consistency rather than absolute agreement. For example, when comparing student performance on the ACT and SAT, a measure of consistency would be helpful to know whether the general ranking, or relative position of students, remains similar despite the ACT and SAT having different scoring scales. If raters are asked to independently rate something, such as observed anti-social behavior, and if raters also develop and use different rating scales, Pearson r could assess whether scores obtained from the two rating scales and raters provided similar relative ratings of those observed for anti-social behavior.

When comparing parallel scales or tests, or when assessing stability of scores from a scale or test, a preferred measure is one that accounts for both relative performance (consistency) and absolute performance (score agreement). Pearson r does not provide a measure that addresses both conditions.

4. Intraclass Correlation Coefficient, ICC

Shrout and Fleiss (1979) introduced three types of ICC, and for two of these types they provided formula for assessing consistency or agreement. For our purposes, we will focus on agreement for situations that likely result from test-retest or parallel forms situations. Shrout and Fleiss called these Case 3 types which use mixed-effects ANOVA to obtain reliability estimates.

The ICC is a measure that partitions variance in scores to account for variation due to respondents, time (in test-retest or forms in parallel forms), and random error.

$$\text{ICC} = \frac{\text{VAR(Respondents)}}{\text{VAR(Respondents)+VAR(Time)+VAR(Error)}} = \frac{\text{VAR(R)}}{\text{VAR(R)+VAR(T)+VAR(E)}}$$

where

VAR(R) = VAR(Respondents) = variance in scores attributed to respondents
VAR(T) = VAR(T) = variance across time of administration (or different forms)
VAR(E) = VAR(Error) = variance that remains in scores after controlling VAR(R) and VAR(T)

The VAR above are estimated from a mixed or random effects ANOVA:

VAR(R) = (Mean Square Between – Mean Square Error) / k, where k is number of testing times
VAR(T) = (Mean Square Time – Mean Square Error) / n, where n is number of respondents
VAR(E) = Mean Square Error

There are several options for running this ANOVA in SPSS. One option using the ANOVA command in SPSS is illustrated below with Time 1 and 2 data from Example 4. This example is for illustrative purposes only and students are not expected to use ANOVA to estimate the ICC. Instead, for this course we will use a more convenient method that is illustrated in the next section.

Figure 1: Data Entry in SPSS for obtain ANOVA Estimates to Calculate ICC

	Scores	Time	Respondents
1	98.00	1.00	1.00
2	87.00	1.00	2.00
3	88.00	1.00	3.00
4	77.00	1.00	4.00
5	72.00	1.00	5.00
6	73.00	1.00	6.00
7	62.00	1.00	7.00
8	63.00	1.00	8.00
9	112.00	2.00	1.00
10	107.00	2.00	2.00
11	108.00	2.00	3.00
12	103.00	2.00	4.00
13	98.00	2.00	5.00
14	93.00	2.00	6.00
15	82.00	2.00	7.00
16	77.00	2.00	8.00

Figure 2: SPSS Commands for ANOVA Results

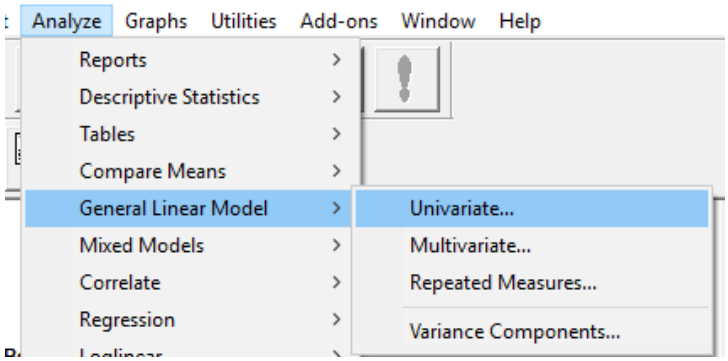


Figure 3: ANOVA Commands in Univariate

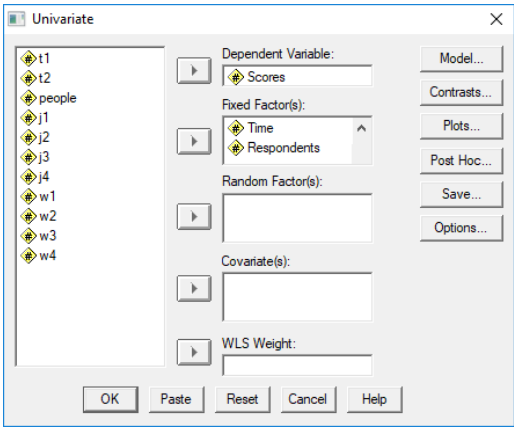


Figure 4: SPSS Univariate ANOVA Output

Tests of Between-Subjects Effects

Dependent Variable: Scores

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	122500.000	1	122500.000	394.797	.000
	Error	2172.000	7	310.286[a]		
Time	Hypothesis	1600.000	1	1600.000	155.556	.000
	Error	72.000	7	10.286[b]		
Respondents	Hypothesis	2172.000	7	310.286	30.[illegible]	.000
	Error	72.000	7	10.286[b]		
Time * Respondents	Hypothesis	72.000	7	10.286	.	.
	Error	.000	0	.[c]		

a. MS(Respondents)
b. MS(Time * Respondents)
c. MS(Error)

MSB
MST
MSE

The three variances needed are:

VAR(R) = (Mean Square Between – Mean Square Error) / k = (310.286-10.286) / 2 = 150.000
VAR(T) = (Mean Square Time – Mean Square Error) / n = (1600-10.286) / 8 = 198.714
VAR(E) = Mean Square Error = 10.286

To calculate ICC for absolute agreement for a single rater or single form, the type we typically see with test-retest and parallel form studies, the formula follows:

$$ICC_{Agreement} = \frac{VAR(R)}{VAR(R)+VAR(T)+VAR(E)} = \frac{150}{150+198.714+10.286} = \frac{150}{359} = .417$$

Recall that the CTT reliability for these data is

$$rxx = \frac{VAR(T)}{VAR(X)} = \frac{140.00}{256.26} = 0.546$$

so the ICC of .417 provides a more realistic estimate of reliability than does the Pearson correlation of .935 provided above.

The ICC can also be calculated to estimate consistency of scores, just like Pearson r. The ICC formula for consistency, rather than agreement, omits variance due to time or differences across the two sets of scores. In this example, the ICC for consistency matches the Pearson r for these data:

$$ICC_{Consistency} = \frac{VAR(R)}{VAR(R) + VAR(E)} = \frac{150}{150+10.286} = \frac{150}{160.286} = .935$$

As stated, the ICC consistency formula above omits the variance for time; this shows that variance resulting from differences in scores across time is ignored, hence the agreement between the ICC for consistency and Pearson r.

5. ICC with SPSS

Below are screenshots showing how to obtain the ICC in SPSS for test-retest and parallel forms reliability. Note that data entry for SPSS requires two columns, one for the first set of scores and one for the second set. It is critical that scores be matched to respondents otherwise obtained estimates will be incorrect; this is also true for Pearson r.

Commands are

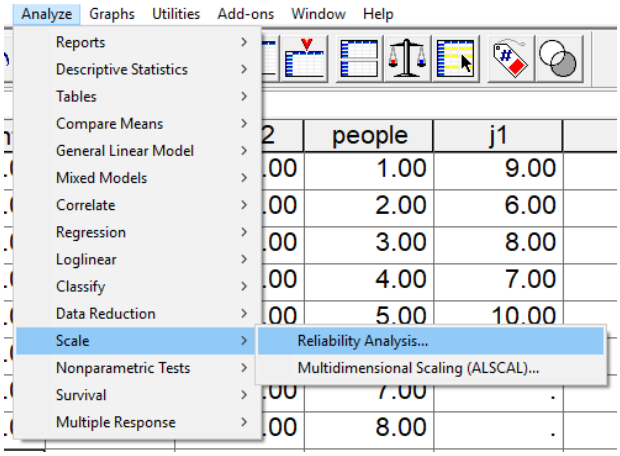
Analyze → Scale → Reliability Analysis

Figure 5: SPSS Data Entry for Test-retest

: Respondent

	Respondent	Time1	Time2
1	1.00	98.00	112.00
2	2.00	87.00	107.00
3	3.00	88.00	108.00
4	4.00	77.00	103.00
5	5.00	72.00	98.00
6	6.00	73.00	93.00
7	7.00	62.00	82.00
8	8.00	63.00	77.00

Figure 6: SPSS Command Selection



With the new screen, identify the two variables that represent the test-retest scores, in this case Time 1 and Time 2 as shown in Figure 7. Move both to the **Items** box.

Figure 7: Selection of Variables

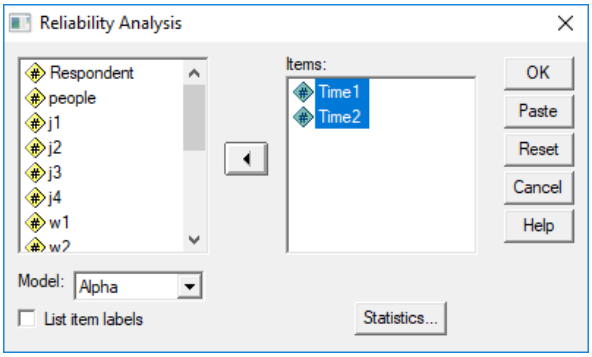


Figure 8: Indicate Which ICC to Calculate

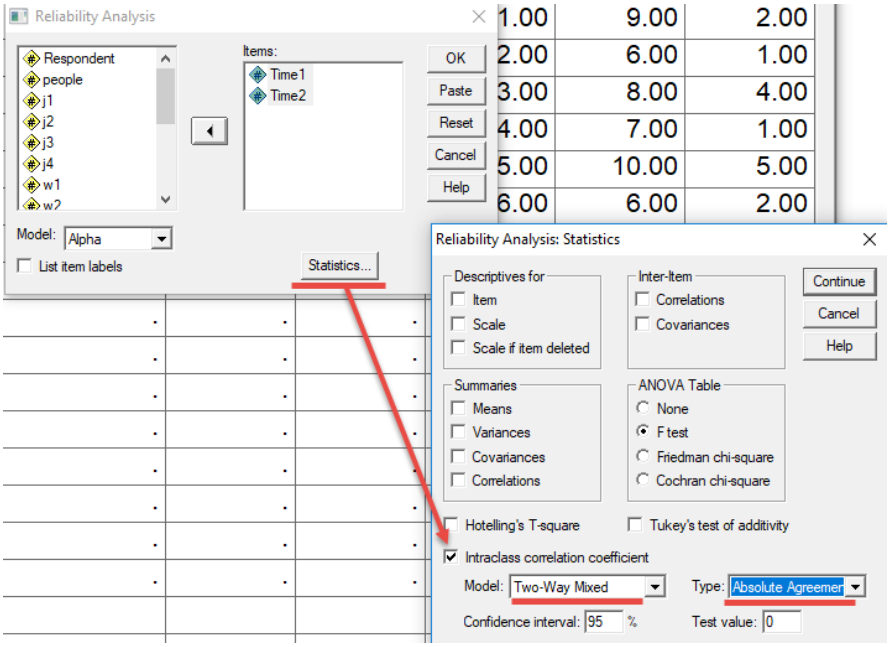


Figure 8 shows that after identifying the two sets of scores, next click on **Statistics**, place a mark next to **Intraclass Correlation Coefficient**, select **Two-way Mixed**, then **Absolute Agreement** as the type.

To obtain estimates, next click **Continue** (Figure 8) and **OK**.

Figure 9: SPSS ICC Results

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.418[b]	-.023	.850	30.167	7.0	7	.000
Average Measures	.589[c]	-.048	.920	30.167	7.0	7	.000

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

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 9 shows tabled results. For test-retest or parallel forms, use the row labeled **Single Measures**, and the ICC for that row is .418. The footnote at the bottom of the table indicates these estimates are from a **two-way mixed model**, and the ICC is based upon **absolute agreement**.

6. ICC Real Data

Students in one of my courses are asked to complete a questionnaire twice. The items were selected from Menon (2001) and were designed to measure three employment related constructs. Responses to each item scaled from Strongly Disagree (1) to Strongly Agree (6). The nine items appear below.

Perceived Control
Q1: I can influence the way work is done in my department
Q2: I can influence decisions taken in my department
Q3: I have the authority to make decisions at work

Goal Internalization
Q4: I am inspired by what we are trying to achieve as an organization
Q5: I am inspired by the goals of the organization
Q6: I am enthusiastic about working toward the organization's objectives

Perceived Competence
Q7: I have the capabilities required to do my job well
Q8: I have the skills and abilities to do my job well
Q9: I have the competence to work effectively

None of these required **reverse coding** (if this term has not been presented already, it will be presented in this course), so composite variables can be computed directly by taken the mean across the three indicators for each construct.

SPSS data file link:

http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/06-reliability/06-EDUR9131-EmploymentThoughts-Merged.sav

Items with _1 are from the first administration, and those with _2 are from the second. Two composite scores have been created for Perceived Control and Goal Internalization using **Transform → Compute** command for Time 1 and 2. (If Transform and Computer for composite scores has not presented already, this will be covered in more detail soon.)

Figure 10: Transform, Compute to Calculate Composite Scores

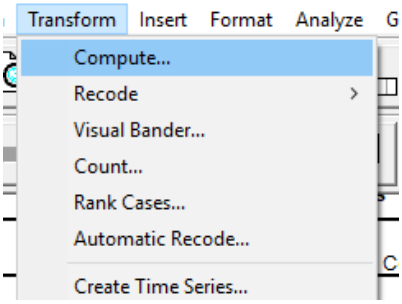
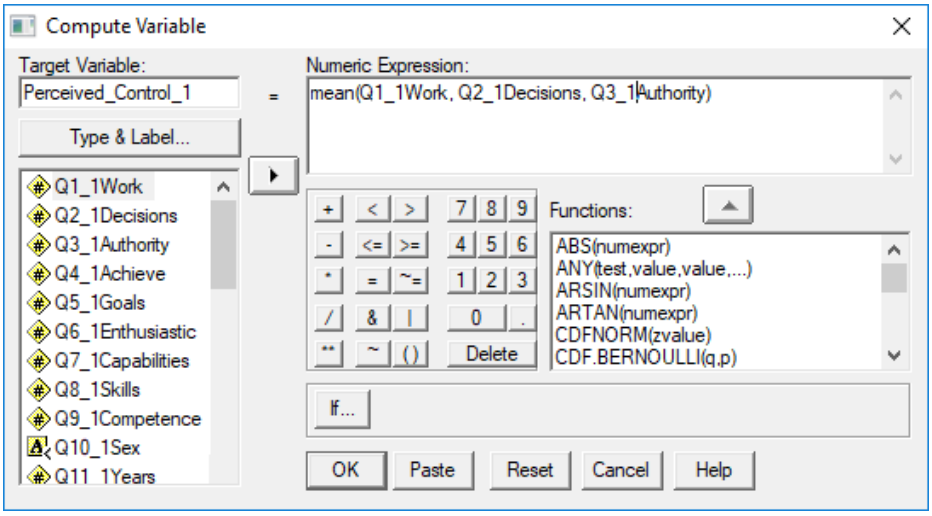


Figure 11: Compute Windows, Use Mean(x,y,z) Function to Calculate Mean Score Composites



Find Pearson r and ICC Absolute Agreement for

- Perceived Control times 1 and 2
- Goal Internalization times 1 and 2

Also

- Compute composite scores for time 1 and 2 for Perceived Competence, then
- Find Pearson r and ICC Absolute Agreement Perceive Competence time 1 and 2

7. Comparison of Results with Menon

Note that estimates of reliability are sample specific so one should always check reliability for each sample. It can be useful to know how well an instrument behaves across samples. How do our results for this sample of graduate students at Georgia Southern compare with results reported by Menon (2001) p. 164?

Menon reports test-retest reliability estimates but does not indicate which form of test-retest estimate was used. Pearson r is the traditional assessment for test-retest reliability, so Menon likely used Pearson r.

For test-retest the following estimates were obtained with the GSU sample data linked above.

		Menon Pearson r?	EDUR 9131 Pearson r	EDUR 9131 ICC Absolute
P. Control	=	.87	.852	.847
Goal Internalization	=	.86	.806	.688
P. Competence	=	.77	.512	.517

8. ICC with More than Two Assessment Periods or Forms

Unlike Pearson r, ICC can be extended to more than two administrations. For example, let us assume that the three measures of Perceived Control

- Q1_1Work
- Q2_1Decisions
- Q3_1Authority

are three administrations of the same scale to the same respondents each a few weeks apart. What would be the ICC, absolute agreement, for these three items?

Figure 12a: SPSS ICC Command for Q1_1Work, Q2_1Decisions, Q3_1Authority

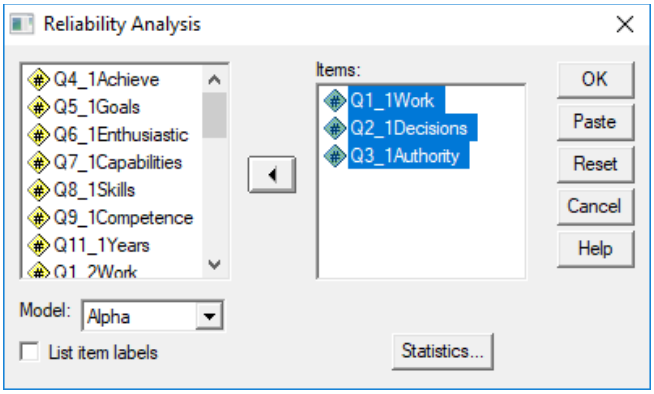


Figure 12b: SPSS ICC Command for Q1_1Work, Q2_1Decisions, Q3_1Authority

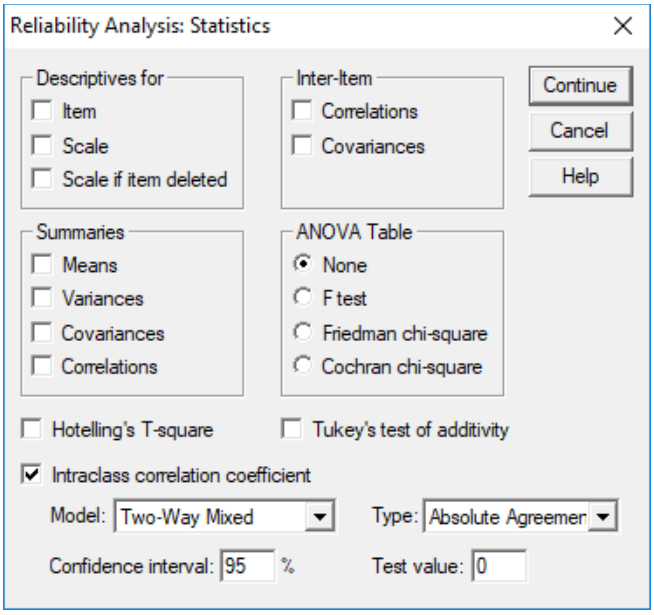


Figure 13: ICC Results for Q1_1Work, Q2_1Decisions, Q3_1Authority

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.788[b]	.603	.906	13.030	17.0	34	.000
Average Measures	.918[c]	.820	.967	13.030	17.0	34	.000

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

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

The value for the ICC of .788 indicates satisfactory level of agreement among the three items. This illustrates that ICC can be extended beyond two time periods for test-retest and beyond two forms with parallel forms.

9. Single Item Test-retest

Sometimes researchers use measures that are composed of just one item. For example, often a single item works well for assessing student evaluations of instruction, e.g.,

Overall, how would you rate this instructor?

Sample response options: Very Poor, Poor, Satisfactory, Good, Very Good

A similar item could be used to measure job satisfaction, e.g.,

Overall, how satisfied are you with your job?

Sample response options: Very Poor, Poor, Satisfactory, Good, Very Good

To assess test-retest reliability for such items, use the same procedures outlined above.

10. Published Examples of Test-retest (to be updated)

Below are sample publications in which test-retest or parallel forms reliability is presented or discussed. These examples help show how to present reliability results.

Published Examples of Test-Retest Assessment

- Kush & Watkins 1996, see Table 2 for test-retest correlations, repeated measures ANOVA to test mean differences p. 317
- Sapountzi-Krepia 2005, see Table 3 for use of Kappa to assess stability of categorical responses; Table 4 for intra-class correlation and rho (Pearson r) for stability and mean differences (also, nice discussion of correlation limitation, alternatives)

Published Examples of Test-Retest Assessment for Single Item

- K Milton, F C Bull, A Bauman 2010 Reliability and validity testing of a single-item physical activity measure, Br J Sports Med; see Tables 3 and 4, Table 1 presents similar studies.
- Moss 2008 Single Item Measures, psychlopedia entry; no tables of examples, instead it presents citations for studies with examples

Qualitative Variables

11. Percent Agreement – See Inter-rater Agree for Nominal Data (presented later in course)
12. Dichotomous Variables: ICC, Kappa, Scott's Pi, Krippendorff's alpha – See Inter-rater Agreement for Nominal Data
13. Ordinal Variables: Weighted Kappa (to be added)

Scoring/rating agreement for judges or evaluators in test-retest situations can be examined using the same procedures for establishing inter-rater agreement. Detail of these procedures are discussed later in the course under agreement for multiple coders or raters when evaluating written responses to open-ended or essay/short answer type items.

For dichotomous (binary) variables, such as pass vs fail, or employed vs not employed, one could use ICC as illustrated above. (More to be added)

References

Bobko, P., Roth, P., & Buster, M. (2007). The usefulness of unit weights in creating composite scores. Organizational Research Methods, 10, 289–709.

Hendrickson A, Patterson B, & Melican G (2008). The Effect of Using Different Weights for Multiple-Choice and Free-Response Item Sections. Presentation at the National Council for Measurement in Education, New York.

Menon, S. (2001). Employee Empowerment: An Integrative Psychological Approach. Applied Psychology: An International Review, 50, 153-180.

Shrout, P.E. and Fleiss, J. L. (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull., 86: 420-428.