

LI Analysis Training Series

Intra-class Correlation for Test-Retest Reliability and/or Stability

(Last Revised: 6/24/99)

Rodney Funk & Michael Dennis

Chestnut Health Systems

Bloomington IL 61701

309-827-6026

www.chestnut.org

Acknowledgement: This document was developed under grant No. T111320 from the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The opinions expressed here are solely those of the authors and do not reflect official positions of the U.S. Government.

Purpose: Standard texts on generalizability theory and psychometrics (e.g., Hayes, 1988; Lord & Novick, 1968; Nunnally & Bernstein, 1994) often recommend testing reliability with an intra-class correlation coefficient. This index represents the percentage of variance explained in the score by individual (or other unit of observation) and ranges from 0 to 1. It can be used to estimate the stability of measure over time (test-retest reliability if a short enough time) and/or to partial-out systematic variation from other sources (e.g., variations in raters, context) that would otherwise be included in the error in a Pearson Product Moment Correlation. Our aim here is to provide an example of how to use SPSS (Version 9.0, 8.0.1 or 7.5) to estimate the intra-class correlation coefficient and its 95% C.I. The first example is from data first appearing in Potthoff and Roy (1964) and is also used in an example in the *SPSS Advanced Statistics 7.5* manual (SPSS, 1997). This example shows how to estimate the intra-class correlation coefficient with four time periods. The second example shows how when calculated on two groups (test-retest), the intra-class correlation coefficient is equal to the Person Moment correlation coefficient, while also showing the difference between reliability and stability.

IMPORTANT: If using 10.0.5, the intra-class correlation coefficient can be calculated using the 'ALPHA' command.

Example 1: The data need to be in a repeated measures format. This data consists of distance in millimeters between the center of the pituitary to the pteryomaxillary fissure, to measure growth. These measurements were collected on 27 adolescents (11 girls and 16 boys). The data originally appeared in Potthoff and Roy (1964). The data is shown in Table 1. Subject was created as a unique identifier, with 1 through 11 assigned to the girls and 12 through 27 to the boys.

Table 1: Sample Data for Example 1 (in 2 columns)

subject	gender	age	distance
1	F	8	21
1	F	10	20
1	F	12	21.5
1	F	14	23
2	F	8	21
2	F	10	21.5
2	F	12	24
2	F	14	25.5
3	F	8	20.5

Subject	Gender	age	distance
3	F	10	24
3	F	12	24.5
3	F	14	26
4	F	8	23.5
4	F	10	24.5
4	F	12	25
4	F	14	26.5
5	F	8	21.5
5	F	10	23

subject	gender	age	distance
5	F	12	22.5
5	F	14	23.5
6	F	8	20
6	F	10	21
6	F	12	21
6	F	14	22.5
7	F	8	21.5
7	F	10	22.5
7	F	12	23
7	F	14	25
8	F	8	23
8	F	10	23
8	F	12	23.5
8	F	14	24
9	F	8	20
9	F	10	21
9	F	12	22
9	F	14	21.5
10	F	8	16.5
10	F	10	19
10	F	12	19
10	F	14	19.5
11	F	8	24.5
11	F	10	25
11	F	12	28
11	F	14	28
12	M	8	26
12	M	10	25
12	M	12	29
12	M	14	31
13	M	8	21.5
13	M	10	22.5
13	M	12	23
13	M	14	26.5
14	M	8	23
14	M	10	22.5
14	M	12	24
14	M	14	27.5
15	M	8	25.5
15	M	10	27.5
15	M	12	26.5
15	M	14	27
16	M	8	20
16	M	10	23.5
16	M	12	22.5

subject	gender	age	distance
16	M	14	26
17	M	8	24.5
17	M	10	25.5
17	M	12	27
17	M	14	28.5
18	M	8	22
18	M	10	22
18	M	12	24.5
18	M	14	26.5
19	M	8	24
19	M	10	21.5
19	M	12	24.5
19	M	14	25.5
20	M	8	23
20	M	10	20.5
20	M	12	31
20	M	14	26
21	M	8	27.5
21	M	10	28
21	M	12	31
21	M	14	31.5
22	M	8	23
22	M	10	23
22	M	12	23.5
22	M	14	25
23	M	8	21.5
23	M	10	23.5
23	M	12	24
23	M	14	28
24	M	8	17
24	M	10	24.5
24	M	12	26
24	M	14	29.5
25	M	8	22.5
25	M	10	25.5
25	M	12	25.5
25	M	14	26
26	M	8	23
26	M	10	24.5
26	M	12	26
26	M	14	30
27	M	8	22
27	M	10	21.5
27	M	12	23.5
27	M	14	25

Procedure : The procedures and equations used here can also be found, along with another example, on pages 169-174 in the *SPSS Advanced Statistics 7.5* manual (SPSS, 1997). To get the desired output in SPSS, choose from the menus:

Statistics
 General Linear Model
 Variance Components

This will open a dialogue box. Using the mouse, click on DISTANCE to highlight it. Then click the arrow next to the box labeled ‘Dependent Variable’ to move the variable to this box. Do the same to move AGE to the ‘Fixed Factor(s)’ box and SUBJECT to the ‘Random Factor(s)’ box. Click on Model. In the Model dialog box, click on Custom and move both AGE and SUBJECT from the Factors & Covariates box to the Method box. Click on Continue. Now click on Options. Under Method, click on Maximum likelihood. Then under Display, click on Iteration history. After clicking on Continue, click the Paste button. The syntax displayed is:

```
VARCOMP
  distance BY subject age
  /RANDOM = subject
  /METHOD = ML
  /CRITERIA = ITERATE(50)
  /CRITERIA = CONVERGE(1.0E-8)
  /PRINT = HISTORY (1)
  /DESIGN = age subject
  /INTERCEPT = INCLUDE .
```

Output: The output needed to calculate the intra-class correlation coefficient and its 95% C.I. are found in the last two items: the Variance estimates for ML (Figure 1) and Asymptotic Covariance matrix for ML (Figure 2). Figure 1 gives the components for calculating the coefficient.

Figure 1.

Variance Estimates	
Component	Estimate
Var(SUBJECT)	4.29944
Var(Error)	2.001486
Dependent Variable: DISTANCE	
Method: Maximum Likelihood Estimation	

The estimate for the intra-class correlation coefficient is:

$$D = \text{Var}(XCID) / (\text{Var}(XCID) + \text{Var}(\text{Error}))$$

For this example: $4.299 / (4.299 + 2.001) = .682$.

For calculating the 95% C.I., it is first necessary to estimate the variance for the correlation coefficient. These numbers are from both Figures 1 and 2.

Figure 2.

Asymptotic Covariance Matrix

	Var(SUBJECT)	Var(Error)
Var(SUBJECT)	1.712714559	-0.02473
Var(Error)	-0.024728064	0.098912

Dependent Variable: DISTANCE

Method: Maximum Likelihood Estimation

Additional Calculations Required. Variance Components below are estimated based on equations in Johnson, Kotz, & Kemp, 1992. From Figure 2, $\text{var}(F^2_s)=1.713$, $\text{var}(F^2_e)=.099$, and $\text{cov}(F^2_s, F^2_e)=-.025$. The equation for estimating the variance for the intra-class correlation coefficient (page 173 in the SPSS manual) is:

$$\begin{aligned} \text{var}(D) &= (\text{var}(\text{Error})^2 * \text{var}(F^2_s) + \text{var}(\text{SUBJECT})^2 * \text{var}(F^2_e) \\ &\quad - 2 * \text{var}(\text{SUBJECT}) * \text{var}(\text{Error}) * \text{cov}(F^2_s, F^2_e)) / (\text{var}(\text{SUBJECT}) + \text{var}(\text{Error}))^4 \\ &= ((2.001)^2 * 1.713 + (4.299)^2 * (.099) \\ &\quad - 2 * 4.299 * 2.001 * (-.025)) / (4.299 + 2.001)^4 \end{aligned}$$

The $\text{var}(D)=.0058$. To get the standard error, take the square root of the variance, which gives a standard error of .076. The 95% C.I. is $\pm 1.96(\text{standard error})$, or $\pm 1.96(.076)=.149$. Thus the 95% C.I. for the first example, $D=.682$, is $(.682 \pm .149)=(.533, .831)$.

Example 2: Again, the data must be in a repeated measures format. There are two rows per client, one for each testing period. The data for the example is given in Table 2. It contains data for 13 adolescents. The variable XCID is the client ID (replaced in the example with arbitrary numbers to protect confidentiality). The test score variable is MRJDYS - the days of marijuana use in the past 90 days. The variable giving the time period is TIME, while XOBSDT gives corresponding test dates.

Table 2: Sample Data for Example 2

Xcid	xobsdt	MRJDYS	time
1	11/6/97	20	1
1	11/21/97	80	2
2	11/19/97	20	1
2	11/21/97	5	2
3	11/25/97	9	1
3	12/12/97	20	2
4	11/26/97	58	1
4	12/12/97	80	2
5	12/8/97	60	1
5	12/12/97	60	2
6	12/4/97	60	1
6	12/19/97	45	2
7	12/5/97	10	2
8	12/12/97	15	1
8	12/19/97	18	2
9	11/24/97	60	1
9	11/28/97	40	2
10	11/3/97	10	1
10	11/21/97	50	2
11	11/4/97	70	1
11	11/21/97	60	2
12	11/19/97	45	1
12	12/5/97	19	2
13	12/17/97	5	1
13	12/19/97	5	2

Procedure : Again, using the menus in SPSS, select the following:

Statistics
 General Linear Model
 Variance Components

Using the mouse, move MRJDYS to the box labeled ‘Dependent Variable’. Move TIME to the ‘Fixed Factor(s)’ box and XCID to the ‘Random Factor(s)’ box. Click on Model. In the Model dialog box, click on Custom and move both TIME and XCID from the Factors & Covariates box to the Method box. Click on Continue. Now click on Options. Under Method, click on Maximum likelihood. Then under Display, click on Iteration history. After clicking on Continue, click the Paste button. If you feel comfortable enough, you can just type the syntax into a syntax window in SPSS. The syntax displayed is:

```
VARCOMP
MRJDYS BY time xcid
/RANDOM = xcid
/METHOD = ML
/CRITERIA = ITERATE(50)
/CRITERIA = CONVERGE(1.0E-8)
/PRINT = HISTORY (1)
/DESIGN = time xcid
/INTERCEPT = INCLUDE .
```

Output: The output needed to calculate the intra-class correlation coefficient and its 95% C.I. are found in Figure 3, the Variance estimates for ML in Figure 4, the Asymptotic Covariance matrix for ML. Figure 1 gives the components for calculating the coefficient.

Figure 3.

Variance Estimates	
Component	Estimate
Var(XCID)	350.2604
Var(Error)	278.142
Dependent Variable: MRJDYS	
Method: Maximum Likelihood Estimation	

The estimate for the intra-class correlation coefficient is:

$$D=350.26/(350.26+278.142)=.557.$$

For calculating the 95% C.I., we first need to estimate the variance for the correlation coefficient. These numbers are from both Figures 3 and 4.

Figure 4.

Asymptotic Covariance Matrix		
	Var(XCID)	Var(Error)
Var(XCID)	39,813.219	-5,950.998
Var(Error)	-5,950.998	11,901.997
Dependent Variable: MRJDYS		
Method: Maximum Likelihood Estimation		

Additional Calculations Required. From Figure 2, $\text{var}(F^2_s)=39813.219$, $\text{var}(F^2_e)=11901.997$, and $\text{cov}(F^2_s, F^2_e)=-5950.998$. Plugging in the appropriate numbers in the equation given in example 1, the equation for estimating the variance for the intra-class correlation coefficient for this example is:

$$\text{var}(D)=\frac{((278.142)^2 * 39813.219 + (350.26)^2 * (11901.997) - 2 * 350.26 * 278.142 * (-5950.998))}{(350.26 + 278.142)^4}$$

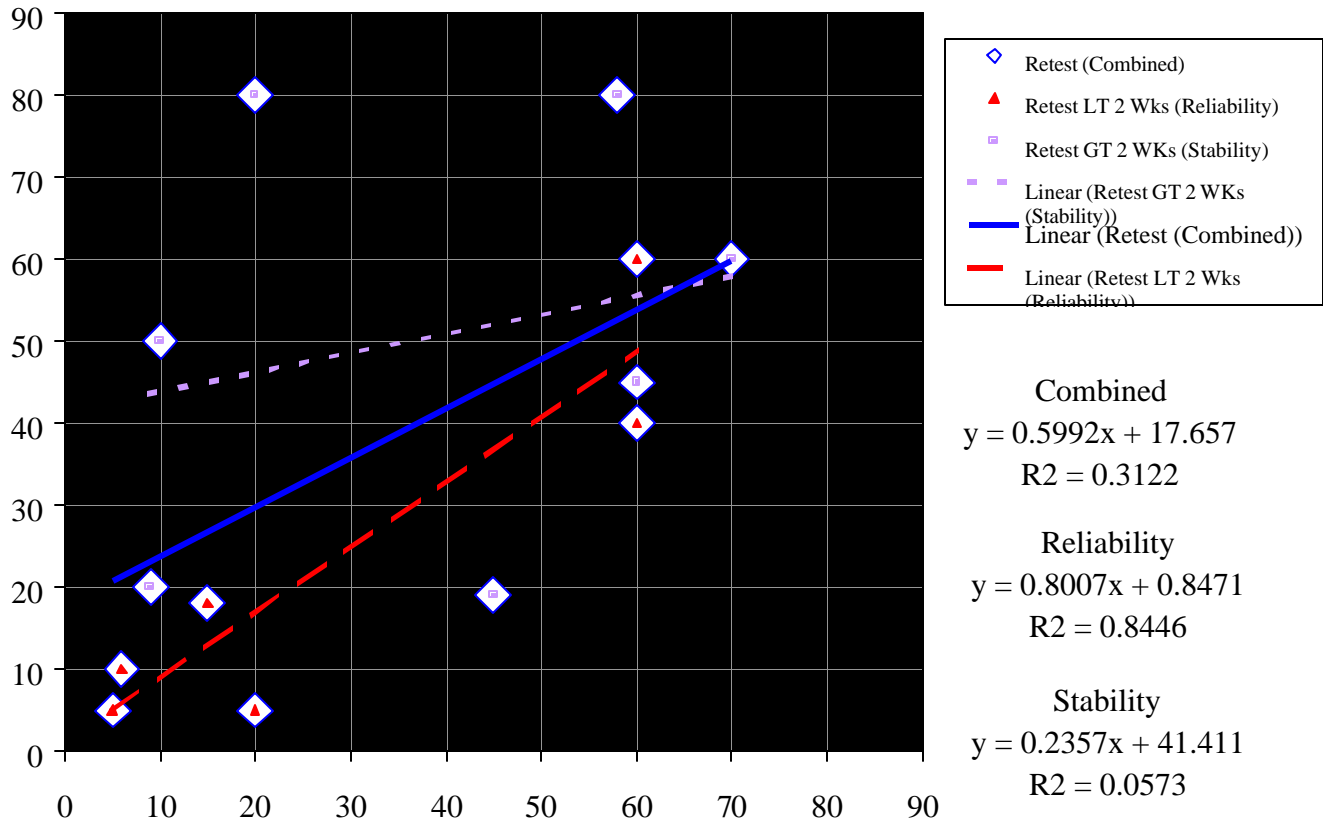
The $\text{var}(D)=.03655$. To get the standard error, take the square root of the variance, which gives a standard error of .119. The 95% C.I. is $\pm 1.96(\text{standard error})$, or $\pm 1.96(.119)=.375$. Thus the 95% C.I. for our example, $D=.557$, is $(.557 \pm .375)=(.182, .932)$. The large confidence interval is due, in part, to having only 13 subjects in this example.

Comments: The intra-class correlation coefficient has been used as a measure of internal consistency, representing the percentage of variance explained by individuals. When there are two measures, as in the second example, the intra-class correlation coefficient and the Pearson moment correlation are almost identical. For our example, the intra-class correlation was .557, while the Pearson Moment correlation is .559. So when only two measures are involved, one can estimate the intra-class correlation coefficient using the Pearson Moment correlation. When there are more than two measures, like in the first example, more detailed calculations are required.

Also note that over short periods of time the intra-class correlation coefficient can be interpreted as reliability. As the amount of time between observations increases, however, the interclass correlation is more appropriately interpreted as a measure of stability (vs. change). The definition of a little or a lot of elapsed time depends on the nature of the measures. As a general rule it should be less than 10% of a time-bound period (e.g., past 90 days, past year) or less than a week. Conversely there should be intervening tasks or at least several hours between administrations to avoid direct recall of the original answers. In the second example, the test to retest time ranged from 2 to 21 days and would actually be more a measure of change/stability than a pure test of reliability. This becomes much clearer when one looks at those re-tested within two weeks or more than two weeks. Figure 5 plots the days reported during the two test periods. A line is plotted for the total, those re-tested within two weeks and those re-tested after two weeks. The corresponding equations and R squares are given. As can be seen, the R^2 for the total sample is .3122. When the re-test was within two weeks, $R^2 = .8446$, while dropping to .0573 when more than two weeks between re-test. So one would conclude that 'days of marijuana use' is a reliable measure, but does not remain stable over time.

Figure 5.

Contrast of Reliability and Stability For Days of Marijuana Use



How important is this? If there are systematic sources of variation these should not be counted as error. Low reliability itself is also one of the major hidden drains on statistical power. The upper limit of an observed correlation ($R_{x'y'}$) is the product of the true correlation (R_{xy}) times the square root of the two reliabilities ($R_{xx} * R_{yy}$) as shown in the following equation:

$$\text{MAXIMUM } R_{x'y'} (\text{observed}) = R_{xy} (\text{true}) * \text{SQRT}(R_{xx} * R_{yy})$$

Thus, the observable correlation will be proportionately reduced by measurement error unless both items are perfectly measured. For example, if both the dependent and independent variable had measurement reliabilities of .7, any “observed” correlations would be reduced by a factor of .7. Table 3 shows the reduction from the “true” correlation to the “observed” correlation for different levels of measurement error in the independent (R_{xx}) and dependent (R_{yy}) variables, as well as their impact on the effect size and required sample sizes. Therefore, if both measures had a reliability of .7, a correlation of .10 would be reduced to .07, .20 would be reduced to .14 and .37 would be reduced to .26. Since smaller “observed” correlations lead to smaller and harder to detect “observed” effect sizes, these reductions are either lost or have to be made up by increasing the corresponding sample sizes.

Table 3. Impact of Measurement Error on R, d and N to Detect Effect (for 80% power, alpha <.05

Max Sqrt(Rxx*Ryy)	Small Effect (*=.20)			Medium Effect (*=.40)			Large Effect (*=.80)		
	Robs	d obs	Min N	Robs	d obs	Min N	Robs	d obs	Min N
Error Free	0.10	0.20	358	0.20	0.40	108	0.37	0.80	29
1.0	0.10	0.20	358	0.20	0.40	108	0.37	0.80	29
0.9	0.09	0.18	484	0.18	0.36	135	0.33	0.71	37
0.8	0.08	0.16	612	0.16	0.32	163	0.30	0.62	46
0.7	0.07	0.14	800	0.14	0.28	216	0.26	0.54	61
0.6	0.06	0.12	1089	0.12	0.24	275	0.22	0.46	83
0.5	0.05	0.10	1569	0.10	0.20	358	0.19	0.38	120
0.4	0.04	0.08	2452	0.08	0.16	612	0.15	0.30	187
0.3	0.03	0.06	4902	0.06	0.12	1089	0.11	0.22	358
0.2	0.02	0.04	9810	0.04	0.08	2452	0.07	0.15	706
0.1	0.01	0.02	39243	0.02	0.04	9810	0.04	0.07	3406

Note: $Robs = \sqrt{Rxx \cdot Ryy} \cdot Rtrue$; $d\ Obs = 2 \cdot Robs / \sqrt{1 - Robs^2}$; where Rxx and Ryy are the reliability of measures x and y.

Source: Dennis, Lennox, & Foss, 1997.

Conversely, one of the easiest and cheapest ways to increase statistical power is to combine highly correlated items into simple indices. Simply switching from the analysis of individual items to sets of related items (which reduces measurement error) can often increase statistical power by 10 to 50% (Dennis, Lennox, & Foss, 1997).

Describing These Procedures. These procedures would normally be described:

We estimated the Interclass Correlation Coefficient (which represents the percentage of variance in the measure explained by individuals) using the Pearson Product Moment correlation coefficient (which is equivalent for 2 observations) as recommended in Hayes (1988) and Winer (1971).

References

- Dennis, M.L., Lennox, R.I., & Foss, M. (1997). Practical power analysis for substance abuse health services research. In K.J Bryant, M Windle, and S.G. West (eds.), The science of prevention: Methodological advances from alcohol and substance abuse research, (Pp. 367-405). Washington, DC: American Psychological Association.
- Hayes, W.L. (1988). Statistics, (Fourth Edition). New York, NY: Holt, Rinehart, and Winston.
- Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). Univariate discrete distributions. New York: John Smiley and Sons.
- Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Nunnally, J.C. & Bernstein, I.H. (1994). Psychometric theory, 3rd edition. New York, NY: McGraw-Hill, Inc.
- Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51, 313-326.
- Statistical Program for the Social Sciences (SPSS 1997). Advanced statistics manual, Version 7.5. Chicago, IL: Author (www.spss.com).

Winer, B.J. (1971). Statistical principles in experimental design. New York, NY: McGraw-Hill Book Company.