Inter-rater Agreement for Ranked Categories of Ratings

1. Ranked Rating Data for Two Raters

If two raters provide ranked ratings, such as on a scale that ranges from strongly disagree to strongly agree or very poor to very good, then Pearson's correlation may be used to assess level of agreement between the raters.

(a) Example—Professional Learning Communities

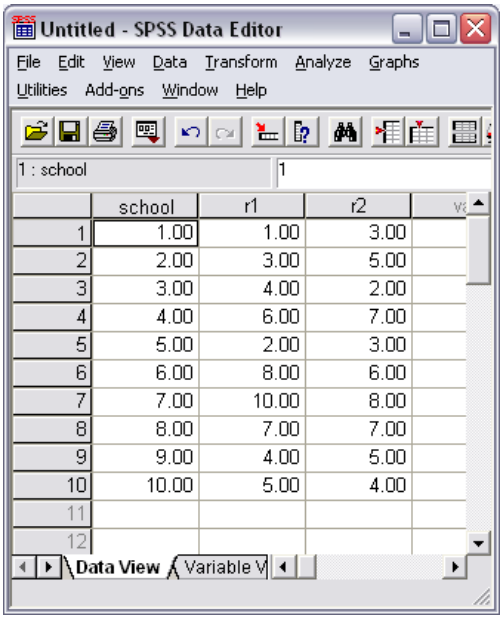
Suppose two raters are asked to rate 10 high schools in terms of level of integration for Professional Learning Communities (PLC). The scale ranges from low of 1 to high of 10

High School	Rater 1	Rater 2
1	1	3
2	3	5
3	4	2
4	6	7
5	2	3
6	8	6
7	10	8
8	7	7
9	4	5
10	5	4

Find the correlation between these two raters. Most researchers design .70 or above, but sometimes .60 is acceptable.

(b) Correlation in SPSS

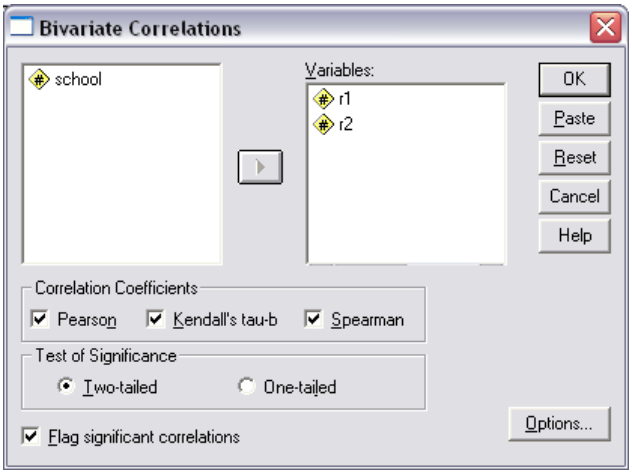
Enter data:



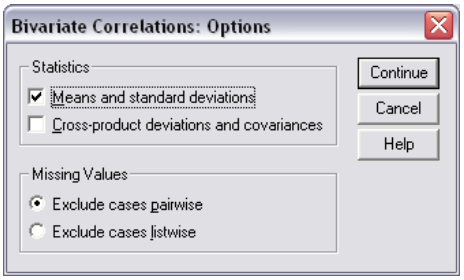
In SPSS, click on

Analyze → Correlate → Bivariate

This opens a pop-up window for correlation. Select the two raters and move both to the variable box. Place marks next to Pearson, Kendall's tau-b, and Spearmen. See below for an example.



Then selection Options and choose Means and Standard Deviations, then select Continue.



Select "OK" to run the correlation.

Results are reported on the next page.

Descriptive Statistics

	Mean	Std. Deviation	N
r1	5.0000	2.78887	10
r2	5.0000	2.00000	10

Correlations

		r1	r2
r1	Pearson Correlation	1	.817(**)
	Sig. (2-tailed)		.004
	N	10	10
r2	Pearson Correlation	.817(**)	1
	Sig. (2-tailed)	.004	
	N	10	10

** Correlation is significant at the 0.01 level (2-tailed).

Correlations

			r1	r2
Kendall's tau_b	r1	Correlation Coefficient	1.000	.628(*)
		Sig. (2-tailed)	.	.014
		N	10	10
	r2	Correlation Coefficient	.628(*)	1.000
		Sig. (2-tailed)	.014	.
		N	10	10
Spearman's rho	r1	Correlation Coefficient	1.000	.801(**)
		Sig. (2-tailed)	.	.005
		N	10	10
	r2	Correlation Coefficient	.801(**)	1.000
		Sig. (2-tailed)	.005	.
		N	10	10

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Some researchers and statisticians argue that Pearson's correlation coefficient is inappropriate when data are just ordinal (ranked data) and therefore should not be used. Alternative correlations for ordinal data include Kendall's tau and Spearman's rho.

- What are the three values for reliability obtained above?
- Is there an acceptable level of agreement between the two raters according to the correlation coefficient?
- Do the mean scores appear to be similar? A correlated samples t-test can be used to assess whether means appear to be similar.

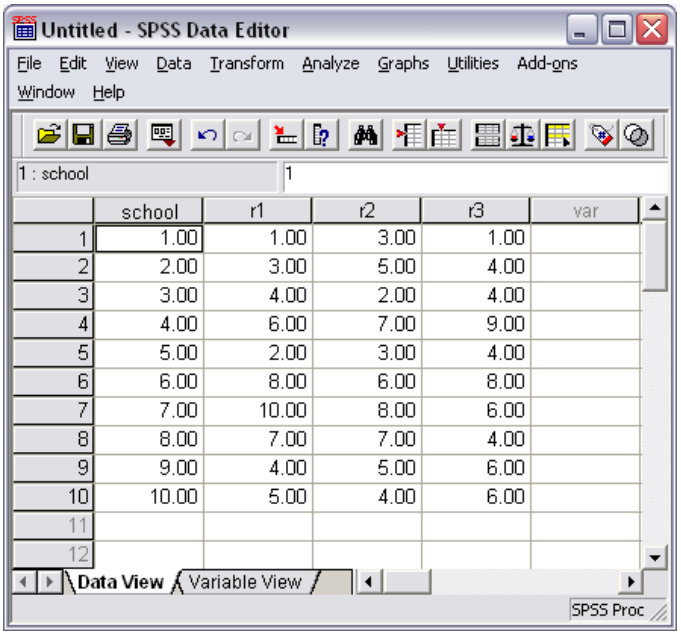
2. Ranked Rating Data for More than Two Raters

(a) Example—Professional Learning Communities

Using the same example as above, suppose we have three raters provide ratings of PLC implementation levels for each school. Below are the data.

High School	Rater 1	Rater 2	Rater 3
1	1	3	1
2	3	5	4
3	4	2	4
4	6	7	9
5	2	3	4
6	8	6	8
7	10	8	6
8	7	7	4
9	4	5	6
10	5	4	6

In SPSS the data appear as illustrated below.

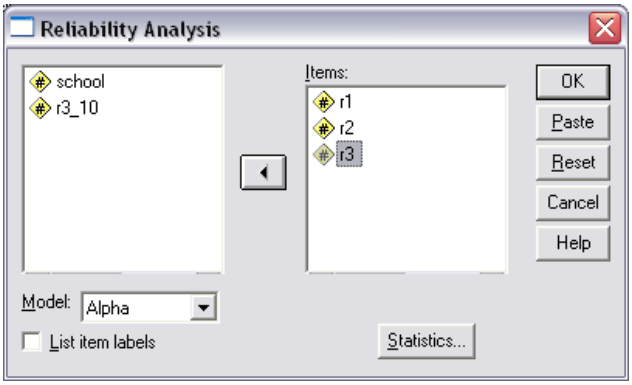


(b) Intra-class Correlation in SPSS

The intra-class correlation is the index most use for determining whether multiple raters using ranked/interval/ratio rating scales provide similar ratings. The commands are:

Analyze → Scale → Reliability Analysis

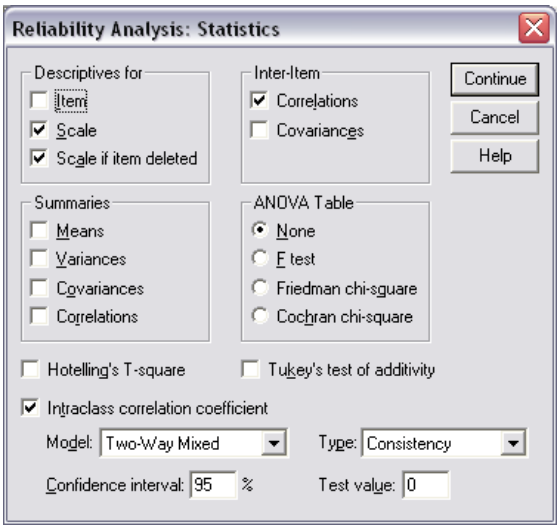
The menu this produces appears below.



Move the raters from the variables box (not labeled) to the “Items” box. Click on “Statistics” and select the following:

- Correlations
- Scale
- Scale if item deleted
- Intraclass correlation coefficient

See image below as illustration.



Click “Continue” then “OK” to obtain results.

Results are reported below.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.856	.864	3

Inter-Item Correlation Matrix

	r1	r2	r3
r1	1.000	.817	.641
r2	.817	1.000	.580
r3	.641	.580	1.000

The covariance matrix is calculated and used in the analysis.

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
r1	10.2000	14.622	.813	.709	.729
r2	10.2000	21.289	.783	.672	.772
r3	10.0000	20.889	.645	.420	.872

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	.664(b)	.313	.892	6.930	9.0	18	.000
Average Measures	.856(c)	.577	.961	6.930	9.0	18	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.
a Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
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 values for the intraclass correlation coefficients are .664 and .856. The first value is the estimate of reliability for a single, typical judge. The second is the estimated reliability if more than one judge is used to provide ratings. Best to report both coefficients.

- What happens if means are different? Multiple rater 3 scores by 10 then rerun the analysis.
- What do the correlation show?
- What do the intraclass correlations show?

(c) Reliability for Ordinal Data

If ordinal data are used, some argue one should use Spearman rho or Kendall tau if there are only two judges or Kendall's coefficient of concordance if there are three or more.

Kendall’s coefficient of concordance to be added.