Assessing the Reliability of Rating Data

Ratings are any kind of coding (qualitative or quantitative) made concerning attitudes, behaviours, or cognitions. Here, I am concerned with those kinds of ratings made by third-parties of a particular individual's attitudes, behaviour, or cognitions. These might be from rating scales, observational check-lists, or symptom check-lists etc. The principle aim of reliability analysis is to determine the degree of agreement between raters when using a particular rating scheme. If the reliability is low, then the scheme itself may be at fault, or the raters, or both! I am not going to try and describe all possible kinds of designs and analyses, but only those that might be most common within the mental health setting.

As always, the quantitative properties of the ratings must be considered first. Then, an appropriate statistic might be chosen to summarise the degree of agreement between raters.

First – an important distinction between inter-rater and intra-class correlations.

Interrater correlation (interrater r).

This is where the similarity between ratings is expressed as a correlation coefficient – generally using a Pearson r product-moment type coefficient. In 2x2 tables (for comparison of just 2 raters), it is possible to use a range of measures of agreement, ranging from the phi coefficient through to say Jaccard's coefficient that excludes all non-occurrences from the calculations). See the DICHOT 3.0 program (downloadable from : **http://www.liv.ac.uk/~pbarrett/programs.htm**) for the implementation of several of these coefficients. For example, working from a 2x2 table with cell IDs as :

	Rater 1 - Yes	Rater 1 - No
Rater 2 - Yes	A	B
Rater 2 - No	C	D

We could compute, for example, the following measures of agreement – each of which takes into account the marginal frequencies in specific ways …

$$Phi = \frac{(A*D - B*C)}{\sqrt{(A+B)*(C+D)*(A+C)*(B+D)}}$$ **the Pearson Product Moment r**

Yule's Q (or gamma) = (A*D - B*C)/(A*D + B*C)

The Jaccard, J = A/(A+B+C)

The G-Index, G = ((A+D)-(B+C))/N

Bennett's B index, B = ((A*D-xkon^2)/((A+xkon)*(D+xkon))
where:

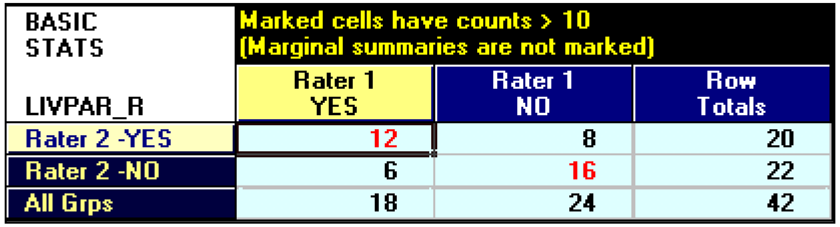
xkon = (B+C)/2
and the Harms and Ihm (1981) adjustment is made (to guard against A or D frequencies = 0)
A = A+1, B= B+1, C= C+1, D=D+1

The output from DICHOT 3.0 shows how they compare (the program includes detailed explanations of the logic of each coefficient in its online help – you already have this as a handout).

Let us take the example where we are looking at the amount of agreement between two raters, on an item from the VRAG

Item 1: Lived with both biological parents to age 16

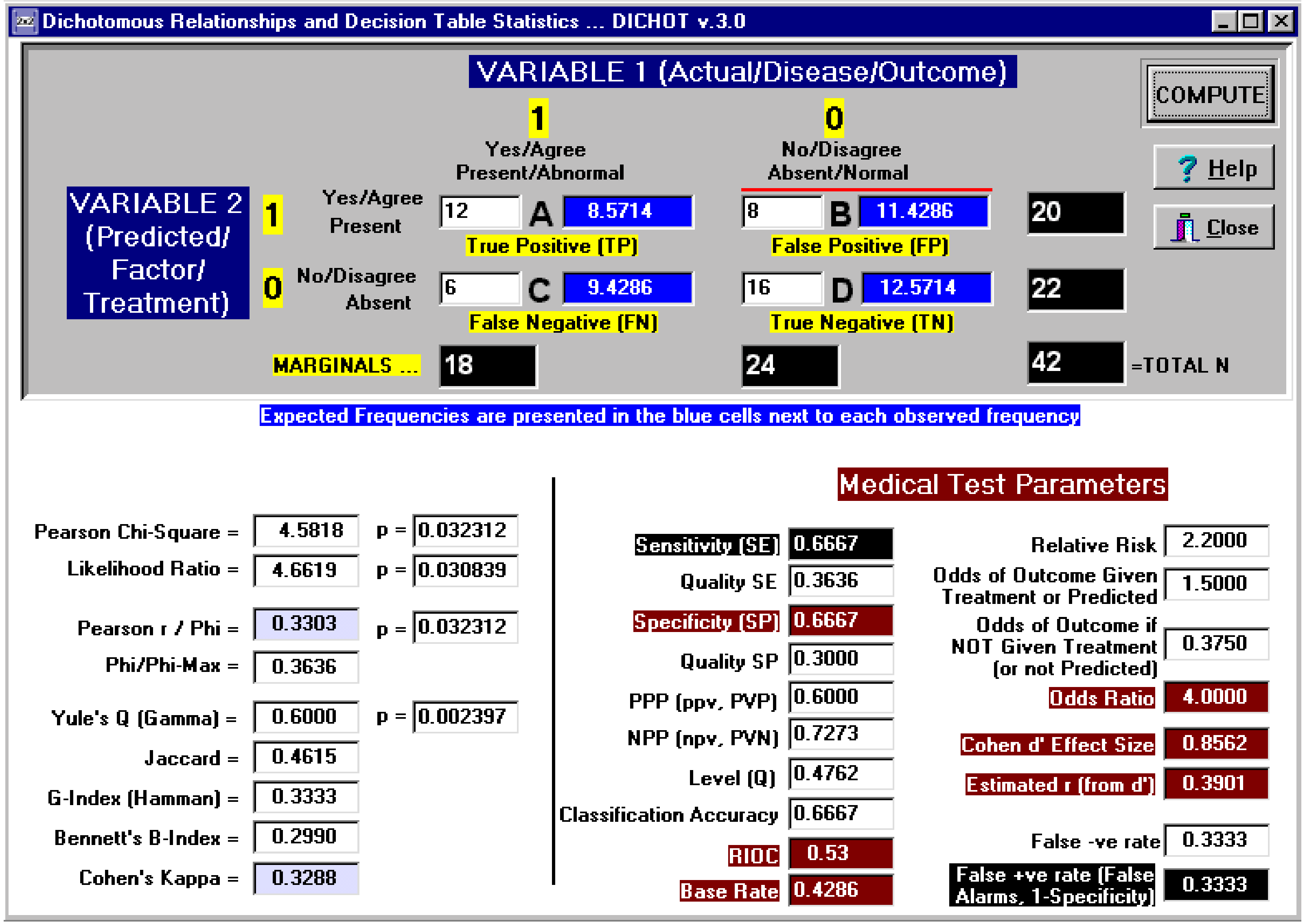
Raw Rating Table



BASIC STATS	Marked cells have counts > 10 (Marginal summaries are not marked)		
LIVPAR_R	Rater 1 YES	Rater 1 NO	Row Totals
Rater 2 -YES	12	8	20
Rater 2 -NO	6	16	22
All Grps	18	24	42

Here we have a simple 2x2 table layout, which we can enter into DICHOT 3.0 for a complete analysis. 42 patients have been rated, and Rater 1 agrees with Rater 2 on (12 + 16) = 28 patients. On 14, there is disagreement.

DICHOT 3.0 Analysis



Dichotomous Relationships and Decision Table Statistics ... DICHOT v.3.0

		VARIABLE 1 (Actual/Disease/Outcome)		
		1 Yes/Agree Present/Abnormal	0 No/Disagree Absent/Normal	
VARIABLE 2 (Predicted/ Factor/ Treatment)	1 Yes/Agree Present	12 A 8.5714 True Positive (TP)	8 B 11.4286 False Positive (FP)	20
	0 No/Disagree Absent	6 C 9.4286 False Negative (FN)	16 D 12.5714 True Negative (TN)	22
	MARGINALS ...	18	24	42 =TOTAL N

Expected Frequencies are presented in the blue cells next to each observed frequency

Pearson Chi-Square =	4.5818	p =	0.032312
Likelihood Ratio =	4.6619	p =	0.030839
Pearson r / Phi =	0.3303	p =	0.032312
Phi/Phi-Max =	0.3636		
Yule's Q (Gamma) =	0.6000	p =	0.002397
Jaccard =	0.4615		
G-Index (Hamman) =	0.3333		
Bennett's B-Index =	0.2990		
Cohen's Kappa =	0.3288		

Medical Test Parameters

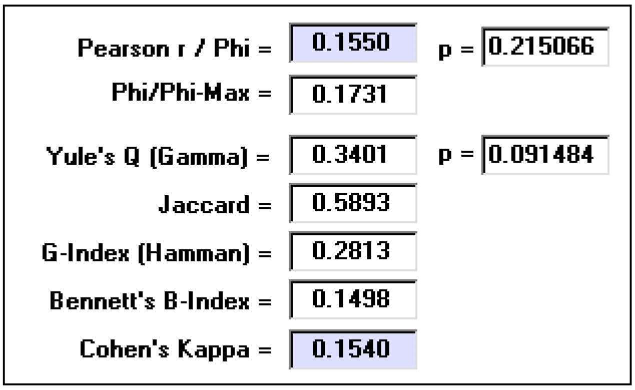
Sensitivity (SE)	0.6667
Quality SE	0.3636
Specificity (SP)	0.6667
Quality SP	0.3000
PPP (ppv, PVP)	0.6000
NPP (npv, PVN)	0.7273
Level (Q)	0.4762
Classification Accuracy	0.6667
RIOC	0.53
Base Rate	0.4286

Relative Risk	2.2000
Odds of Outcome Given Treatment or Predicted	1.5000
Odds of Outcome if NOT Given Treatment (or not Predicted)	0.3750
Odds Ratio	4.0000
Cohen d' Effect Size	0.8562
Estimated r (from d')	0.3901
False -ve rate	0.3333
False +ve rate (False Alarms, 1-Specificity)	0.3333

As can be seen, there is considerable variance between the values of the various coefficients. This is mild compared to some differences that may be observed. What is important is that you understand the rationale behind the coefficient being used, and are thus able to interpret its value accordingly. Play around with DICHOT 3.0 to see just how far the values can sometimes vary. For example, take the table below…

	Rater 1 - Yes	**Rater 1 - No**
Rater 2 - Yes	**33**	**13**
Rater 2 - No	**10**	**8**

Where 64 patients are rated on a Yes/No rating variable. They agree on *Yes* for 33 patients, and on *No* for 8, the remaining patients are classified differentially by the raters. The results:



Pearson r / Phi = 0.1550 p = 0.215066
Phi/Phi-Max = 0.1731
Yule's Q (Gamma) = 0.3401 p = 0.091484
Jaccard = 0.5893
G-Index (Hamman) = 0.2813
Bennett's B-Index = 0.1498
Cohen's Kappa = 0.1540

Kappa agreement (Cohen's Kappa)

Kappa was designed specifically as a measure of agreement between 2 judges, **where ratings are categorical, and where a correction for *chance agreement* is made.** This coefficient thus differs from the percent agreement approach adopted by some, because this simple calculation does not take into account what the chance-level agreement between judges would be alone, assuming they both guessed randomly. The formula for kappa computed for any number of ratings categories used by two raters/judges is:

$$\kappa = \frac{\sum f_o - \sum f_e}{N - \sum f_e} \qquad \text{where} \sum f_o = \text{observed frequencies in the diagonal}$$

$$\sum f_e = \text{expected frequencies in the diagonal}$$

$$N = \text{Number of Patients}$$

The expected frequencies are the same as those calculated for the Pearson Chi-Square calculation, except we use just the diagonal values (A and D) for both observed and expected frequencies. In contrast to this formula, we might consider use of the Jaccard coefficient , which is another measure of interrater agreement, but one that excludes joint-negatives from its calculation. A useful point is that both kappa and the Jaccard coefficients can be interpreted as % values. Kappa can be interpreted as the % agreement after correcting for chance. The Jaccard coefficient can be interpreted as the % agreement after excluding joint negative pairs. Both coefficients vary between **0** and **1** (or **0** to **100%**). DICHOT 3.0 computes this coefficient for 2x2 tables.

If we extend our example to an analysis of the reliability of a 3-point rating, we might have as an example…

	Rater 1 - High	Rater 1 - Med	Rater 1- Low
Rater 2 - High	**5**	**3**	**4**
Rater 2 - Med	**0**	**7**	**3**
Rater 2 - Low	**0**	**0**	**3**

Here we have 25 patients rated by two raters, using a high-medium-low rating frame. The diagonal expected frequencies generated under a hypothesis of independence are:

	Rater 1 - High	Rater 1 - Med	Rater 1- Low
Rater 2 - High	**2.4**		
Rater 2 - Med		**4**	
Rater 2 - Low			**1.2**

Our formula is:

$$\kappa = \frac{\sum f_o - \sum f_e}{N - \sum f_e} \qquad \text{where} \sum f_o = \text{observed frequencies in the diagonal}$$

$$\sum f_e = \text{expected frequencies in the diagonal}$$

$$N = \text{Number of Patients}$$

So…..

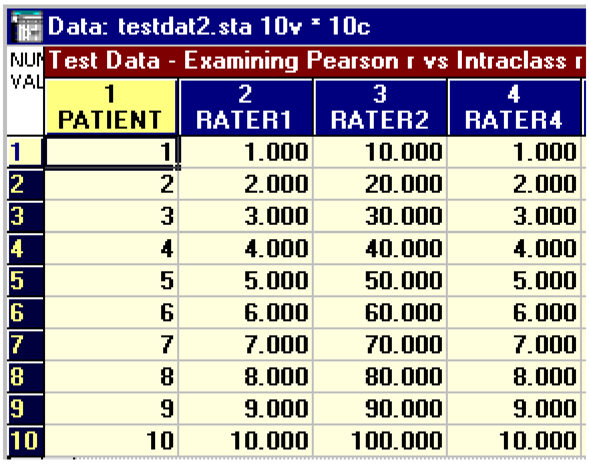
$$\kappa = \frac{\sum f_o - \sum f_e}{N - \sum f_e} = \frac{(5+7+3)\text{-}(2.4+4+1.2)}{25\text{-}(2.4+4+1.2)} = \frac{15-7.6}{25-7.6} = 0.43$$

kappa for these data = 0.43.

Intraclass Correlation (Intraclass r)

This coefficient corrects for a fatal flaw with interrater correlation computed using product-moment correlations. **That is, interrater r takes no account of the variance between the raters.** Remember that product-moment correlations use standardized data, which effectively removes the component of individual rater variability. Essentially, product moment correlations are insensitive to scale, but sensitive to monotonicity relations between data. A simple example to how misleading interrater correlations can be is given below:

Artificial Data file – 10 patients, 3 raters (100 point rating scale)



Data: testdat2.sta 10v * 10c

NUM VAL Test Data - Examining Pearson r vs Intraclass r

	1 PATIENT	2 RATER1	3 RATER2	4 RATER4
1	1	1.000	10.000	1.000
2	2	2.000	20.000	2.000
3	3	3.000	30.000	3.000
4	4	4.000	40.000	4.000
5	5	5.000	50.000	5.000
6	6	6.000	60.000	6.000
7	7	7.000	70.000	7.000
8	8	8.000	80.000	8.000
9	9	9.000	90.000	9.000
10	10	10.000	100.000	10.000

Computing the interrater r (**pearson correlation**) between raters 1 and 2, we get **1.00** **(even though the ratings differ drastically)**

The **Intraclass r** (Shrout and Fleiss model 2) assumes that each patient is rated by two or more raters. These raters are randomly selected from a larger population of raters. Each rater rates all patients. *(In effect, a two-way ANOVA random effects model)* is **0.056.**

Computing the interrater r (**pearson correlation**) between raters 1 and 4, we also get **1.00** (now the ratings truly are identical). The **Intraclass r** for these data is also **1.00**

This simple example indicates why the **intraclass** r
is always to be preferred to interrater r.

Before we delve into the computations and compute-file layouts for three types of intraclass correlation (the Shrout and Fleiss models 1, 2, and 3), it is worthwhile to mention two other methods of assessing interrater reliability. For interval-level data, we might **use coefficient alpha**, and if our ratings are to be considered ordinal, we would use **Kendall's Coefficient of Concordance** (I have provided the relevant pages from Siegel and Castellan's textbook for Kendall's coefficient). When using the alpha coefficient, we are making a measure of the *internal consistency* between raters. It is in fact algebraically equivalent to the ***intraclass correlation*** coefficient where there is only one rating (dependent) variable (or item) being rated and **IF we assume that the judges' ratings are to be averaged to produce a**

<u>composite rating</u>. Essentially, this coefficient tells you how reliable that ratings are <u>as a whole</u> (how internally consistent are the judges' ratings). However, because of this "averaging" of ratings, we reduce the variability of the judges ratings such that when we average all judges ratings, we effectively remove all the error variance for judges.

Take a look at the ANOVA formula below …

$$r_{ic}^2 = \frac{MS_p - MS_r}{MS_p + ((n_j - n_{jav}) \cdot MS_r)/n_{jav}}$$

where

MS_p = mean square effect for Patients/Persons

MS_r = mean square residual effect

n_j = the number of raters/judges

n_{jav} = the numbers of judges to be averaged

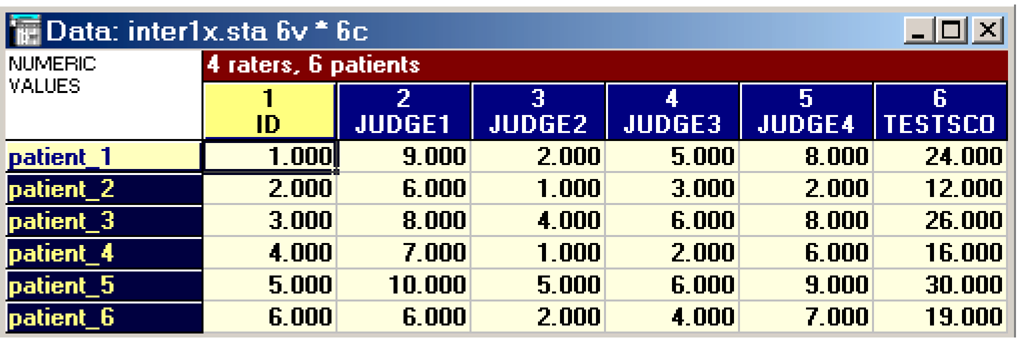
Now, when $n_j = n_{jav}$ we have…

$$r_{ic}^2 = \frac{MS_p - MS_r}{MS_p + ((n_j - n_{jav}) \cdot MS_r)/n_{jav}} = \frac{MS_p - MS_r}{MS_p + ((n) \cdot MS_r)/n_{jav}}$$

$$r_{ic}^2 = \frac{MS_p - MS_r}{MS_p}$$

which in fact is an alternative formula for coefficient alpha, the measure of internal-consistency that we are familiar with in questionnaire psychometrics.

Out of interest, let's look at a problem where we compute our interrater reliability using coefficient alpha. The data file looks like:



Data: inter1x.sta 6v * 6c

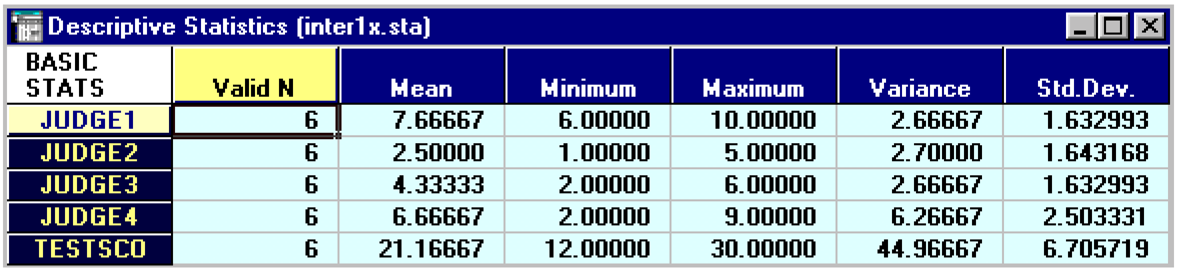
NUMERIC VALUES	4 raters, 6 patients					
	1 ID	2 JUDGE1	3 JUDGE2	4 JUDGE3	5 JUDGE4	6 TESTSCO
patient_1	1.000	9.000	2.000	5.000	8.000	24.000
patient_2	2.000	6.000	1.000	3.000	2.000	12.000
patient_3	3.000	8.000	4.000	6.000	8.000	26.000
patient_4	4.000	7.000	1.000	2.000	6.000	16.000
patient_5	5.000	10.000	5.000	6.000	9.000	30.000
patient_6	6.000	6.000	2.000	4.000	7.000	19.000

Each patient is rated by a judge, on a 1-10 point rating scale. Assuming the data are equal-interval, we compute **coefficient alpha as 0.909.** In essence, we have treated the judges as "items" in a questionnaire, and our patients are the "observations" on these items. Thus, we are in effect doing an "item analysis".

The conventional (one that might look familiar to you!) formula for alpha we are using is:

$$\alpha = \frac{k}{k-1} \cdot \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{S_T^2} \right)$$

where k = the number of items (judges)
s_i^2 = item (judge) variance i of k
S_T^2 = the total test score variance
and…



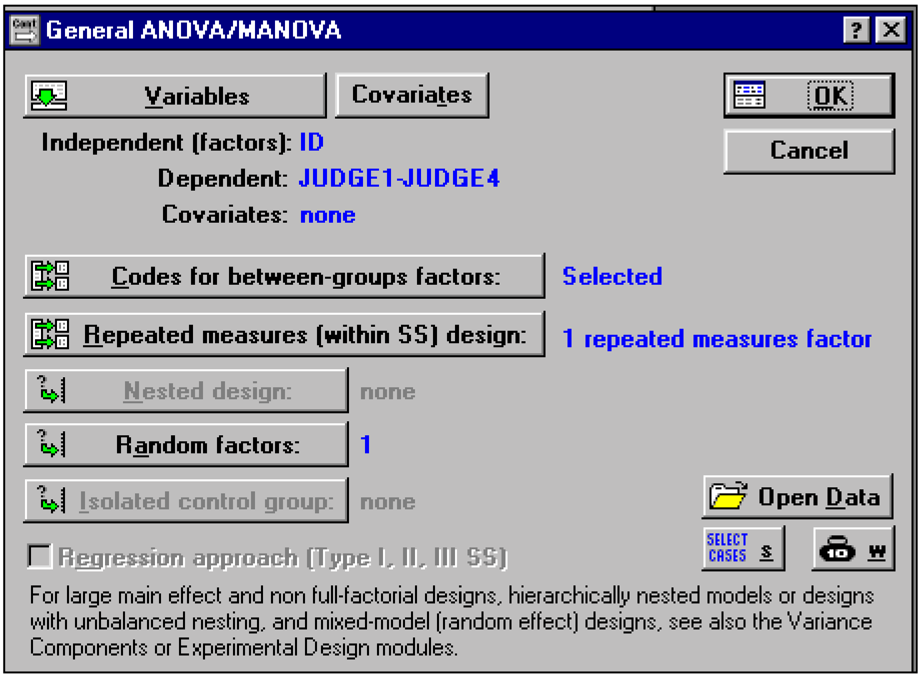
Descriptive Statistics (inter1x.sta)

BASIC STATS	Valid N	Mean	Minimum	Maximum	Variance	Std.Dev.
JUDGE1	6	7.66667	6.00000	10.00000	2.66667	1.632993
JUDGE2	6	2.50000	1.00000	5.00000	2.70000	1.643168
JUDGE3	6	4.33333	2.00000	6.00000	2.66667	1.632993
JUDGE4	6	6.66667	2.00000	9.00000	6.26667	2.503331
TESTSCO	6	21.16667	12.00000	30.00000	44.96667	6.705719

$$\alpha = \frac{k}{k-1} \cdot \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{S_T^2} \right) = \frac{4}{3} \cdot \left(1 - \left(\frac{2.66667 + 2.7 + 2.66667 + 6.26667}{44.96667} \right) \right) = 0.909$$

If we compute a 2-way ANOVA on the data file, with Judges as the repeated measures factor, we obtain …

Statistica ANOVA setup screen, with Patients as random effects



General ANOVA/MANOVA

Independent (factors): ID

Dependent: JUDGE1-JUDGE4

Covariates: none

Codes for between-groups factors: Selected

Repeated measures (within SS) design: 1 repeated measures factor

Nested design: none

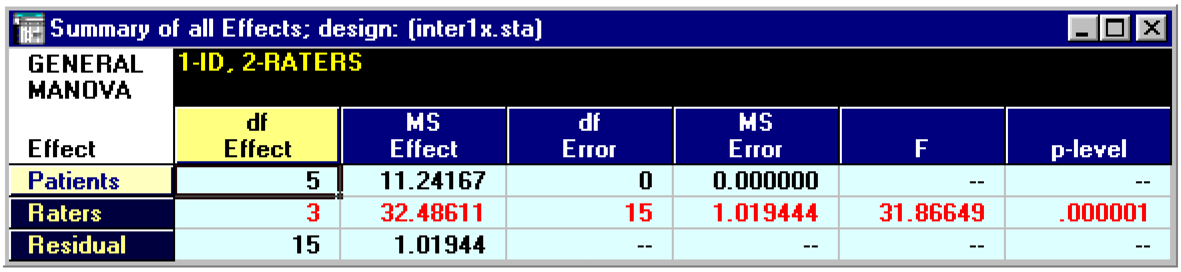
Random factors: 1

Isolated control group: none

□ Regression approach (Type I, II, III SS)

For large main effect and non full-factorial designs, hierarchically nested models or designs with unbalanced nesting, and mixed-model (random effect) designs, see also the Variance Components or Experimental Design modules.

And …



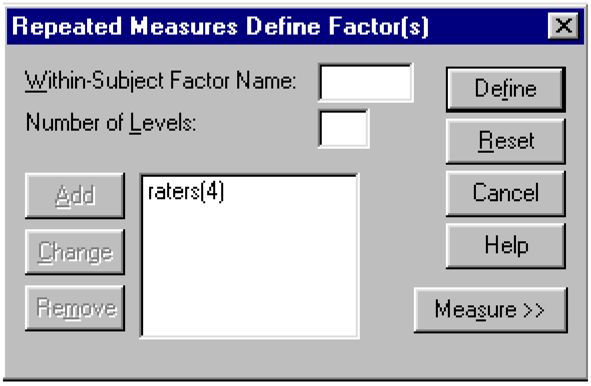
Summary of all Effects; design: (inter1x.sta)

GENERAL MANOVA Effect	1-ID, 2-RATERS df Effect	MS Effect	df Error	MS Error	F	p-level
Patients	5	11.24167	0	0.000000	--	--
Raters	3	32.48611	15	1.019444	31.86649	.000001
Residual	15	1.01944	--	--	--	--

Which, if we now use the ANOVA formula for alpha gives us …

$$r_{ic}^2 = \frac{MS_p - MS_r}{MS_p} = \alpha = \frac{11.24167 - 1.01944}{11.24167} = 0.909$$

The SPSS 9/10 commands to generate these data are via the **Analyze** Menu, then **General Linear Model**, with submenu "**Repeated Measures**". Then setup the Raters factor …



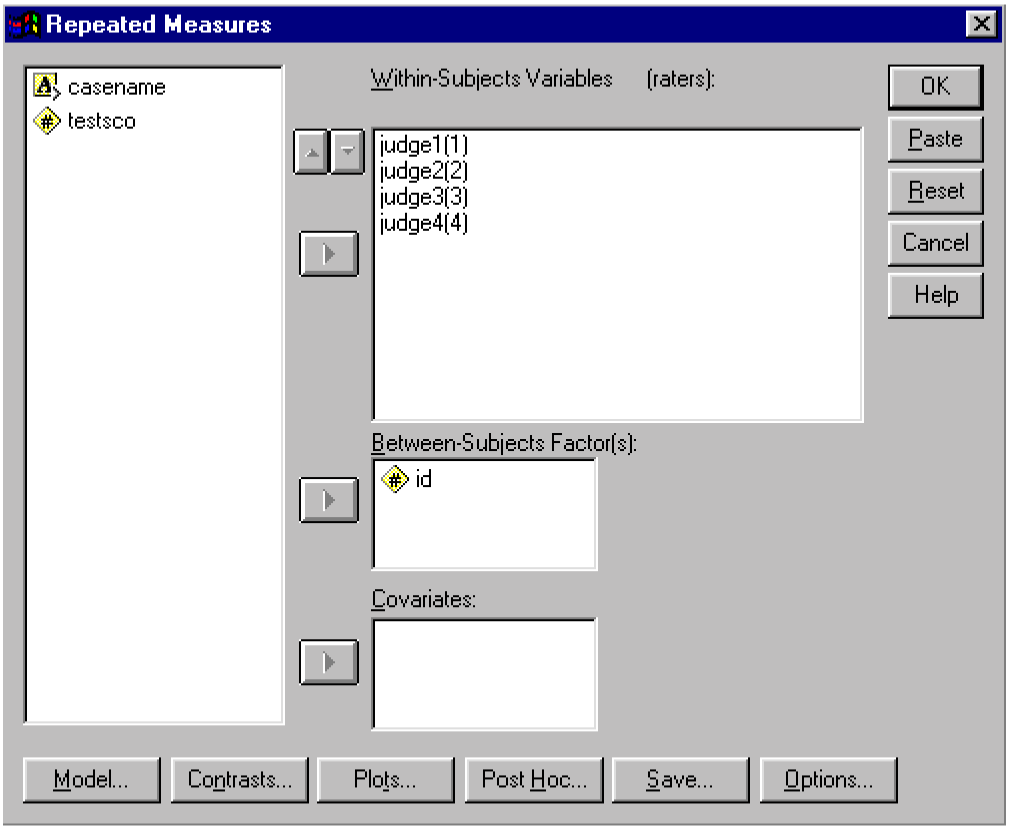
Repeated Measures Define Factor(s)

Within-Subject Factor Name:

Number of Levels:

raters(4)

Press Define and make the selections so as to look like this…



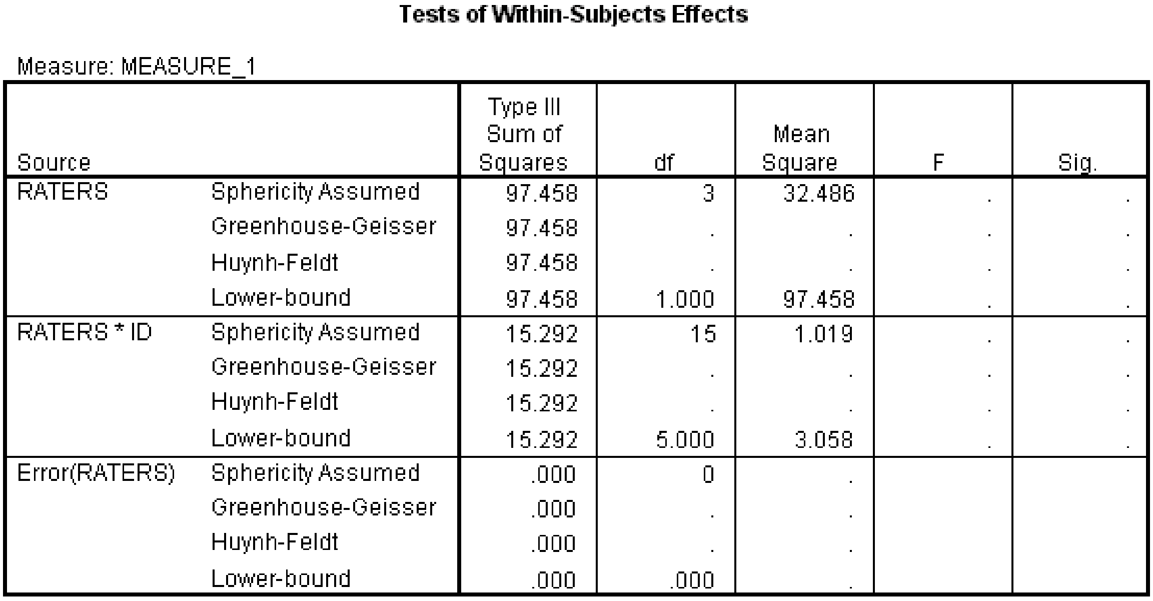
casename
testsco

Within-Subjects Variables (raters):
judge1(1)
judge2(2)
judge3(3)
judge4(4)

Between-Subjects Factor(s):
id

Covariates:

Then … OK … and these are the results …

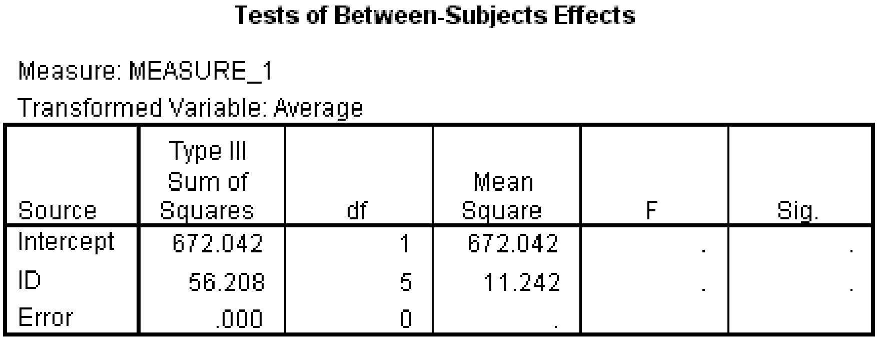


Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
RATERS	Sphericity Assumed	97.458	3	32.486	.	.
	Greenhouse-Geisser	97.458	.	.	.	.
	Huynh-Feldt	97.458	.	.	.	.
	Lower-bound	97.458	1.000	97.458	.	.
RATERS * ID	Sphericity Assumed	15.292	15	1.019	.	.
	Greenhouse-Geisser	15.292	.	.	.	.
	Huynh-Feldt	15.292	.	.	.	.
	Lower-bound	15.292	5.000	3.058	.	.
Error(RATERS)	Sphericity Assumed	.000	0	.		
	Greenhouse-Geisser	.000	.	.		
	Huynh-Feldt	.000	.	.		
	Lower-bound	.000	.000	.		

And …



Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	672.042	1	672.042	.	.
ID	56.208	5	11.242	.	.
Error	.000	0	.		

Anyway, after that digression, let’s go back to the three main designs that encompass Intraclass Correlation reliability designs.

Here, I am following the treatment outlined in the excellent chapter by Orwin (1996) who reports the seminal work by Shrout and Fleiss (1979). Some of the below can also be easily recast within generalizability theory approaches (see Crocker and Algina (1986) – but this is so confusingly demonstrated that I much prefer the clarity of Orwin and Shrout and Fleiss.

Essentially, there are three models that concern us:

Model 1: Each patient to be rated is rated by a unique rater, with each rater randomly selected from a larger population (a one-way ANOVA random effects model). Specifically, for every patient variable or item to be rated, there is a unique rater. Each rater makes only one rating decision. This model assumes you have a large pool of raters, who are randomly assigned to make one rating per patient per variable. So, for a study in which we rate 10 patients on 5 variables, we would need 50 raters. The ANOVA formula is:

$$r_1^2 = \frac{MS_p - WMS}{MS_p + (n_r - 1) * WMS}$$

where $MS_p =$ Between Patients mean square

$n_r =$ number of raters and $n_p =$ number of patients

$MS_{res} =$ Residual mean square

$WMS =$ Within Patients mean square

$MS_r =$ Between Raters ("measures") mean square

with

$$WMS = \frac{\left[\left(MS_r * (n_r - 1)\right) + \left(MS_{res} * (n_p - 1) * (n_r - 1)\right)\right]}{n_p * (n_r - 1)}$$

Model 2: Every patient is rated by each rater. We assume the raters are randomly selected from some population of raters (a two-way random effects model). In essence, each rater rates all patients on all variables. This is the default model that covers most rating situations. For example, for a study in which we rate 10 patients on 5 variables, we would need at least 2 raters in order to assess interrater reliability. Each rater would make (10*5)=50 rating judgements. The ANOVA formula is:

$$r_2^2 = \frac{MS_p - MS_{res}}{MS_p + (n_r - 1) * MS_{res} + \left(\frac{n_r * \left(MS_r - MS_{res}\right)}{n_p}\right)}$$

where $MS_p =$ Between Patients mean square

$MS_{res} =$ Residual (interaction) mean square

Model 3: Every patient is rated by each rater, **<u>BUT</u>**, in contrast to Model 2, **we assume the raters are THE population of raters** (a two-way, fixed-effects model). In essence, each rater rates all patients on all variables. For example, for a study in which we rate 10 patients on 5 variables, we would select say 2 raters in order to assess interrater reliability. Each rater would make (10*5)=50 rating judgements. **However, it is assumed that these are the only two raters who will ever make ratings – no generalizability assumed to other raters.** The ANOVA formula is:

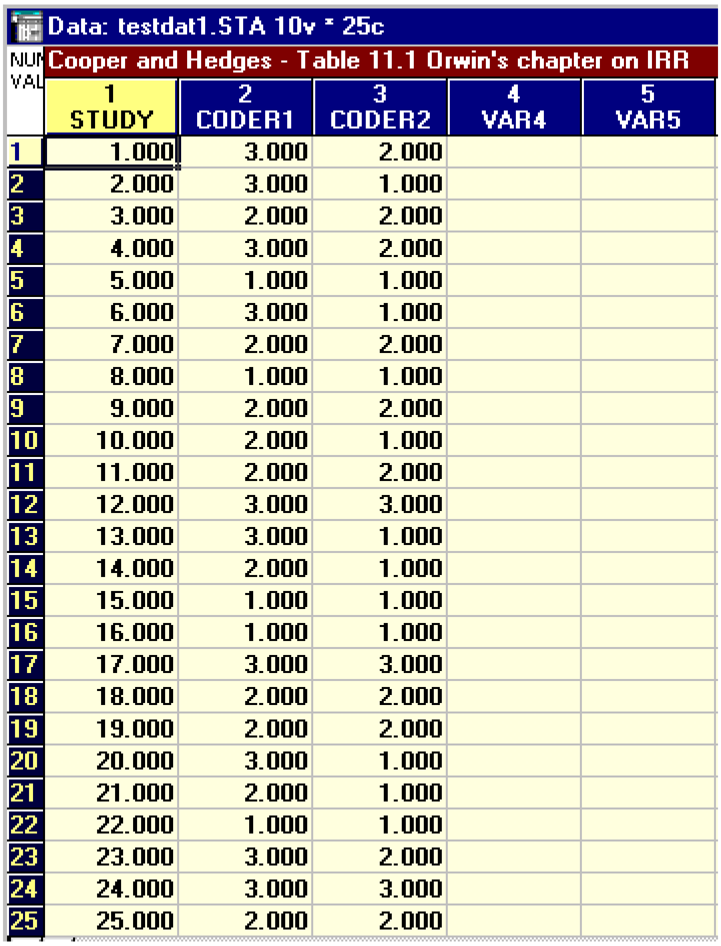
$$r_3^2 = \frac{MS_p - MS_{res}}{MS_p + (n_r - 1) * MS_{res}}$$

$$\text{where } MS_p = \text{Between Patients mean square}$$

$$n_r = \text{number of raters}$$

$$MS_{res} = \text{Residual mean square}$$

Let us take an example dataset from Orwin (1994) …



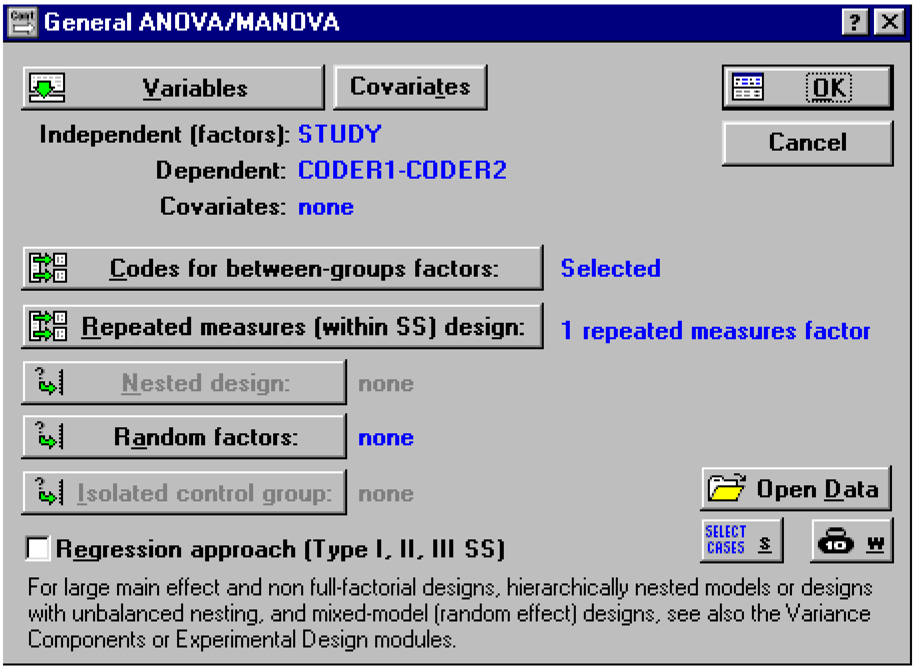
Data: testdat1.STA 10v * 25c

NUI VAL Cooper and Hedges - Table 11.1 Orwin's chapter on IRR

	1 STUDY	2 CODER1	3 CODER2	4 VAR4	5 VAR5
1	1.000	3.000	2.000		
2	2.000	3.000	1.000		
3	3.000	2.000	2.000		
4	4.000	3.000	2.000		
5	5.000	1.000	1.000		
6	6.000	3.000	1.000		
7	7.000	2.000	2.000		
8	8.000	1.000	1.000		
9	9.000	2.000	2.000		
10	10.000	2.000	1.000		
11	11.000	2.000	2.000		
12	12.000	3.000	3.000		
13	13.000	3.000	1.000		
14	14.000	2.000	1.000		
15	15.000	1.000	1.000		
16	16.000	1.000	1.000		
17	17.000	3.000	3.000		
18	18.000	2.000	2.000		
19	19.000	2.000	2.000		
20	20.000	3.000	1.000		
21	21.000	2.000	1.000		
22	22.000	1.000	1.000		
23	23.000	3.000	2.000		
24	24.000	3.000	3.000		
25	25.000	2.000	2.000		

Where we have ratings made on the quality of 25 studies on a 3-point rating scale.

In Statistica, the ANOVA results for these data … with this setup:



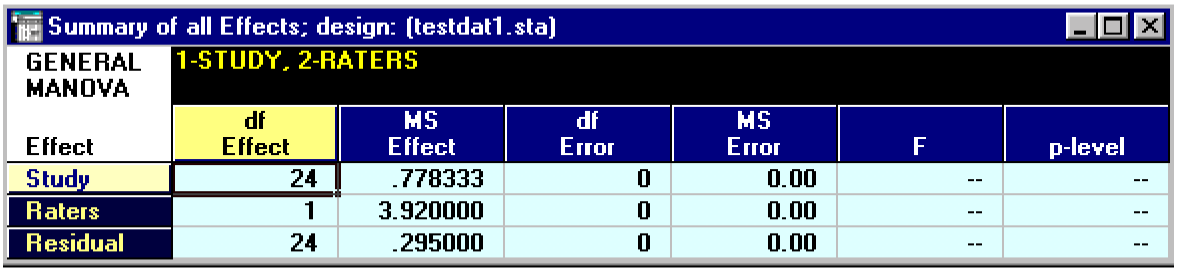
Independent (factors): STUDY
Dependent: CODER1-CODER2
Covariates: none

Codes for between-groups factors: Selected
Repeated measures (within SS) design: 1 repeated measures factor
Nested design: none
Random factors: none
Isolated control group: none

☐ Regression approach (Type I, II, III SS)

For large main effect and non full-factorial designs, hierarchically nested models or designs with unbalanced nesting, and mixed-model (random effect) designs, see also the Variance Components or Experimental Design modules.

Are:



GENERAL MANOVA — 1-STUDY, 2-RATERS

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
Study	24	.778333	0	0.00	--	--
Raters	1	3.920000	0	0.00	--	--
Residual	24	.295000	0	0.00	--	--

If we assumed that each rating for each study was given by a unique rater (random raters), we have Model 1 intraclass r

$$r_1^2 = \frac{MS_p - WMS}{MS_p + (n_r - 1) * WMS} \quad \text{where } MS_p \text{ now } = \text{ studies being rated } (n_p = 25)$$

$$WMS = \frac{\left[\left(MS_r * (n_r - 1)\right) + \left(MS_{res} * (n_p - 1) * (n_r - 1)\right)\right]}{n_p * (n_r - 1)}$$

$$WMS = \frac{((3.92 * 1) + (0.295 * 24 * 1))}{25 * 1} = 0.44$$

$$r_1^2 = \frac{0.778333 - 0.44}{0.778333 + (1) * 0.44} = 0.28$$

If we assume that two raters (assumed to be a sample from some population of raters) provided ratings of each of the 25 studies, then we have Model 2 intraclass r:

$$r_2^2 = \frac{MS_p - MS_{res}}{MS_p + (n_r - 1) * MS_{res} + \left(\frac{n_r * (MS_r - MS_{res})}{n_p} \right)}$$

$$r_2^2 = \frac{0.778333 - 0.295}{0.778333 + (1) * 0.295 + \left(\frac{2 * (3.92 - 0.295)}{25} \right)} = 0.354$$

However, if we assumed that the raters were the only ones we could ever use, essentially the population of raters, then we have Model 3 intraclass r =

$$r_3^2 = \frac{MS_p - MS_{res}}{MS_p + (n_r - 1) * MS_{res}}$$

$$r_3^2 = \frac{0.778333 - 0.295}{0.778333 + (1) * 0.295} = 0.45$$

Our three **Intraclass** r’s are:	Model 1 = 0.28
	Model 2 = 0.35
	Model 3 = 0.45

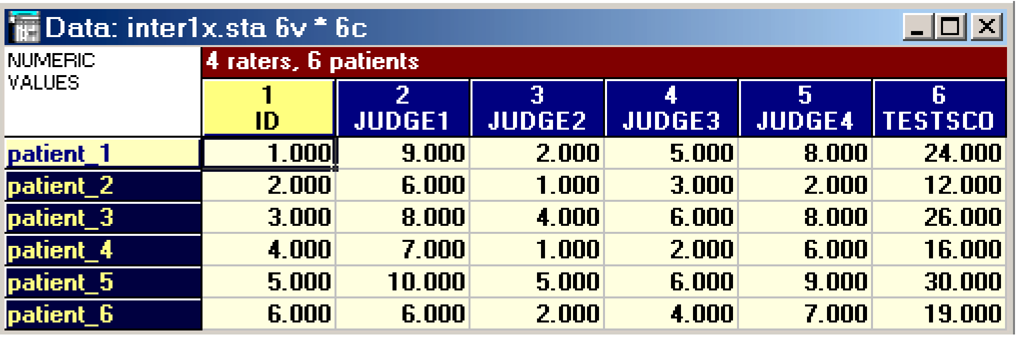
The example on page 5 is actually these data transformed into a table suitable for Kappa – where we assumed the ratings were categorical . The value computed was:

Kappa = 0.43

A **Pearson r** correlation for the same data = 0.45

Kendall’s W (coefficient of Concordance …. 0.40
assuming ordinal categories

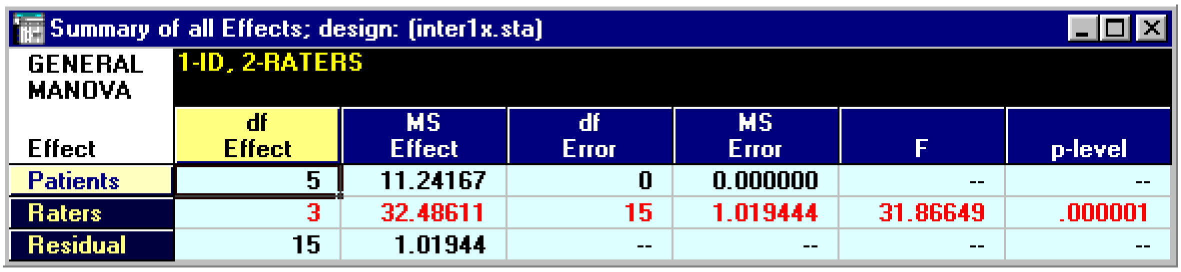
So, returning to our 4 judges data …



Data: inter1x.sta 6v * 6c

NUMERIC VALUES	4 raters, 6 patients					
	1 ID	2 JUDGE1	3 JUDGE2	4 JUDGE3	5 JUDGE4	6 TESTSCO
patient_1	1.000	9.000	2.000	5.000	8.000	24.000
patient_2	2.000	6.000	1.000	3.000	2.000	12.000
patient_3	3.000	8.000	4.000	6.000	8.000	26.000
patient_4	4.000	7.000	1.000	2.000	6.000	16.000
patient_5	5.000	10.000	5.000	6.000	9.000	30.000
patient_6	6.000	6.000	2.000	4.000	7.000	19.000

With ANOVA results as:

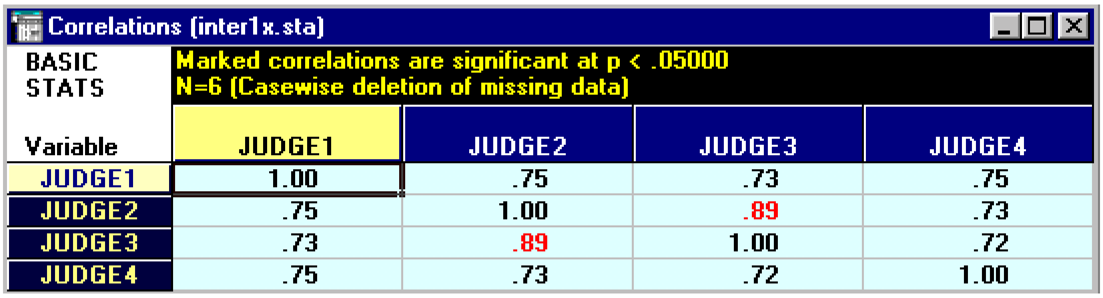


Summary of all Effects; design: (inter1x.sta)

GENERAL MANOVA	1-ID, 2-RATERS					
Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
Patients	5	11.24167	0	0.000000	--	--
Raters	3	32.48611	15	1.019444	31.86649	.000001
Residual	15	1.01944	--	--	--	--

Our three **Intraclass** r’s are: Model 1 = 0.17
Model 2 = 0.29
Model 3 = 0.71

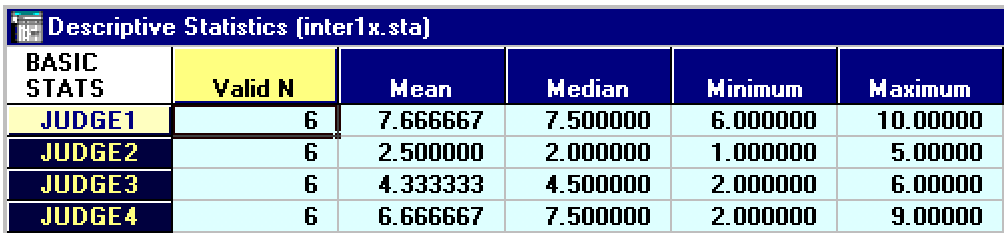
The Mean Inter-Judge Pearson correlation = 0.76



Correlations (inter1x.sta)

BASIC STATS	Marked correlations are significant at p < .05000 N=6 (Casewise deletion of missing data)			
Variable	JUDGE1	JUDGE2	JUDGE3	JUDGE4
JUDGE1	1.00	.75	.73	.75
JUDGE2	.75	1.00	.89	.73
JUDGE3	.73	.89	1.00	.72
JUDGE4	.75	.73	.72	1.00

And now look at how the judges have used their rating scales …



Descriptive Statistics (inter1x.sta)

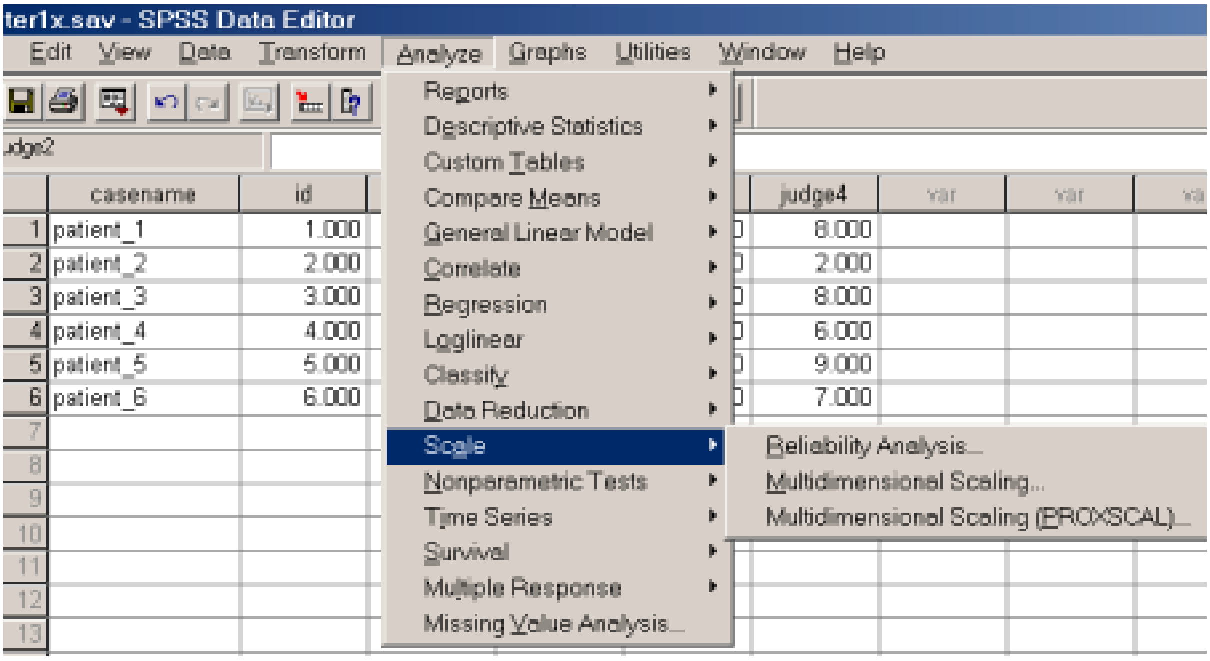
BASIC STATS	Valid N	Mean	Median	Minimum	Maximum
JUDGE1	6	7.666667	7.500000	6.000000	10.00000
JUDGE2	6	2.500000	2.000000	1.000000	5.00000
JUDGE3	6	4.333333	4.500000	2.000000	6.00000
JUDGE4	6	6.666667	7.500000	2.000000	9.00000

SPSS Windows v.9/10 GUI examples for all three models

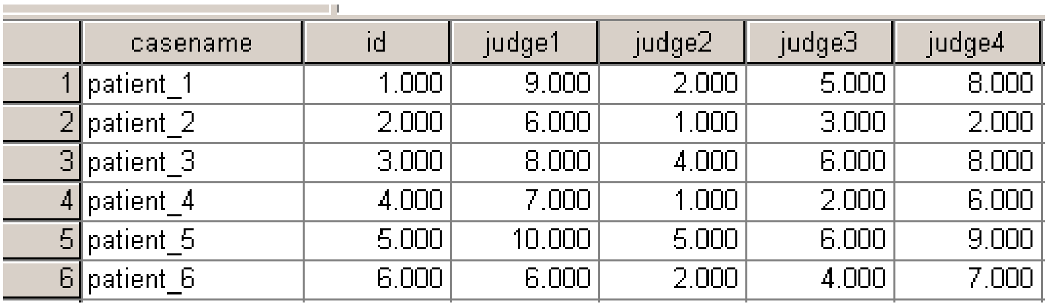
It is instructive to compare the terminology and use of SPSS 9/10 to compute the Intraclass coefficients for Models 1, 2, and 3 above.

“**People Effects**” in the SPSS dialogs equate to **Patients** in my dialog.
“Item Effects” in the SPSS dialogs equate to **Raters** in my dialog

SPSS can directly compute the Intraclass correlation using the **Reliability** option from the **Scale** option on the **Analyze** main menu.



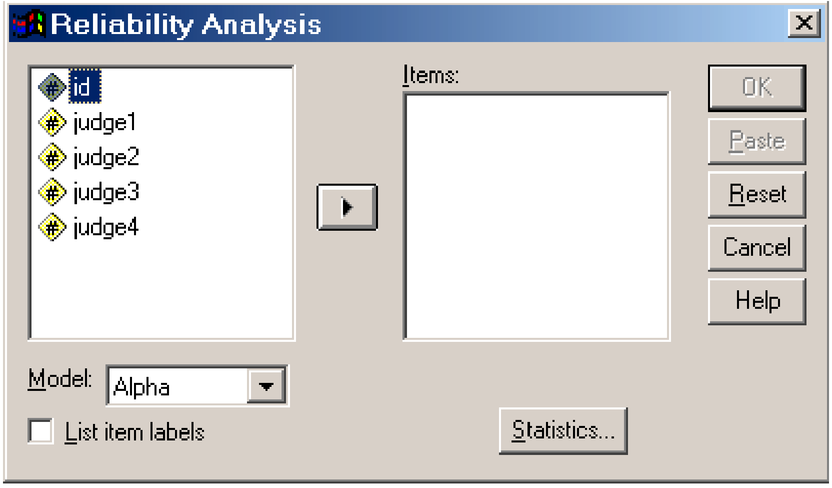
Using the 6 patient x 4 judges dataset as before …



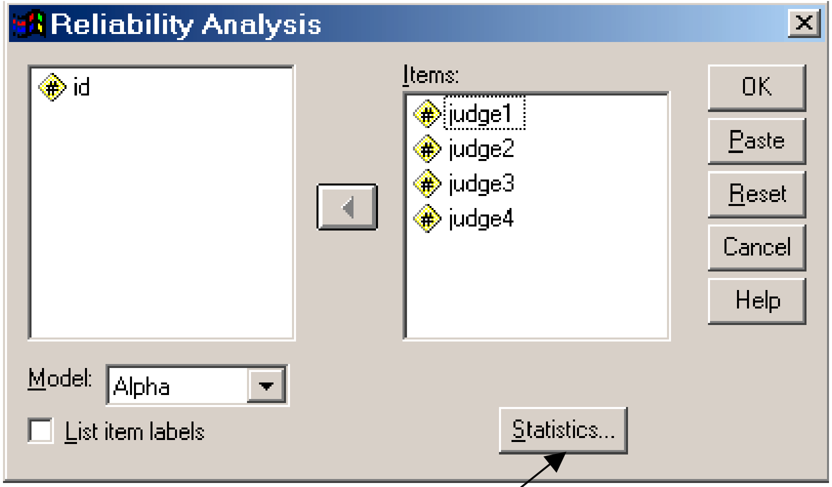
	casename	id	judge1	judge2	judge3	judge4
1	patient_1	1.000	9.000	2.000	5.000	8.000
2	patient_2	2.000	6.000	1.000	3.000	2.000
3	patient_3	3.000	8.000	4.000	6.000	8.000
4	patient_4	4.000	7.000	1.000	2.000	6.000
5	patient_5	5.000	10.000	5.000	6.000	9.000
6	patient_6	6.000	6.000	2.000	4.000	7.000

Model 1: Each patient to be rated is rated by a unique rater, with each rater randomly selected from a larger population (a one-way ANOVA random effects model). Specifically, for every patient variable or item to be rated, there is a unique rater. Each rater makes only one rating decision. This model assumes you have a large pool of raters, who are randomly assigned to make one rating per patient per variable. So, for a study in which we rate 10 patients on 5 variables, we would need 50 raters.

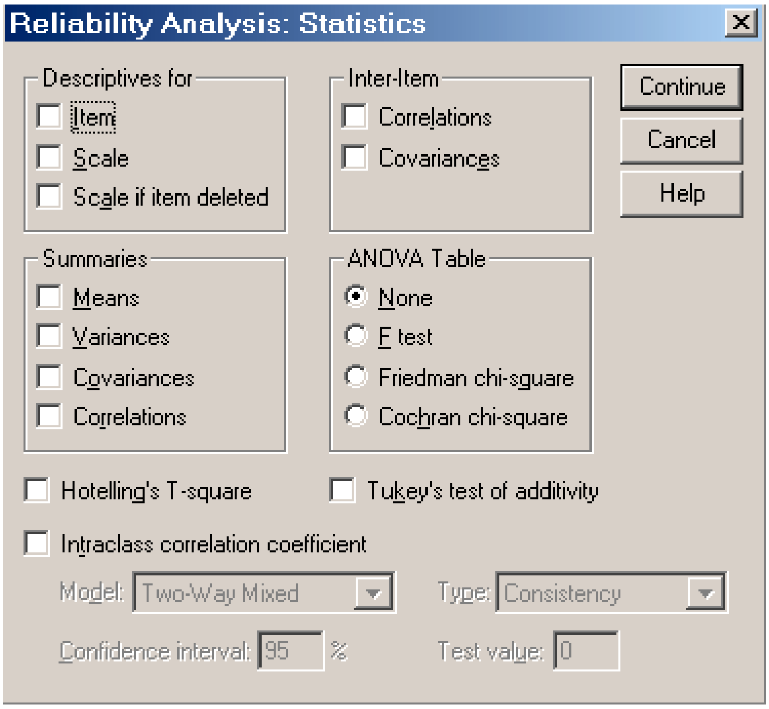
The Reliability analysis screen looks like …



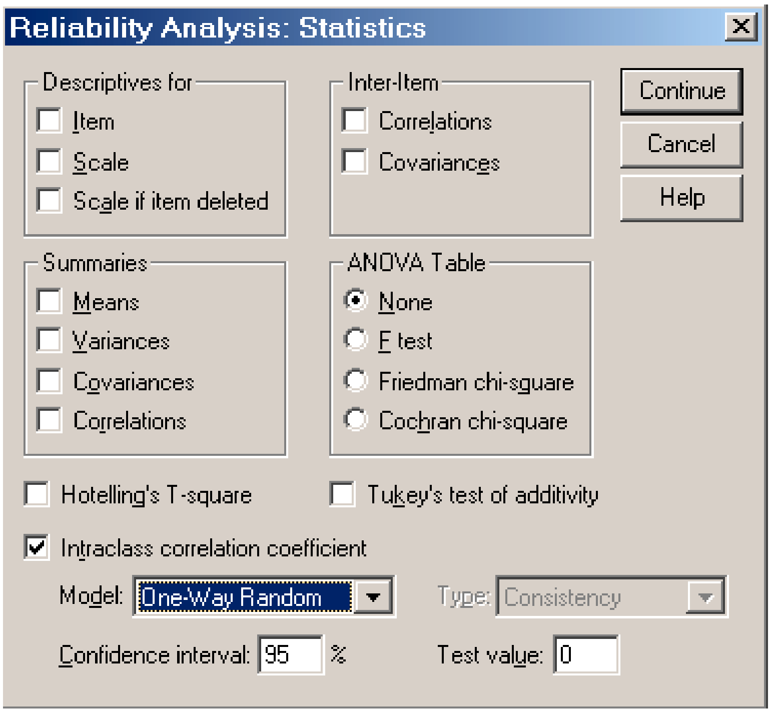
Select the 4 judges as “items” … with Model = Alpha



Then click the “Statistics” button



Then select Intraclass correlation coefficient and One Way Random Model (note that the “type” box is greyed out)….



☑ Intraclass correlation coefficient

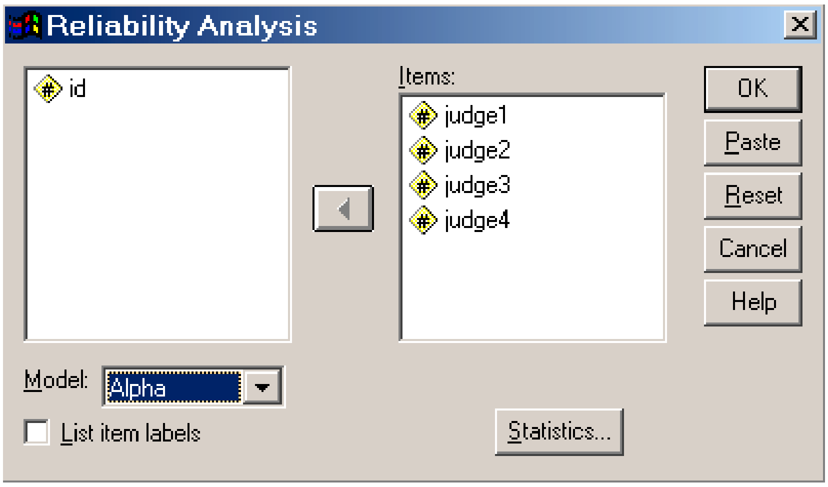
Model: One-Way Random

Type: Consistency

Confidence interval: 95 %

Test value: 0

Then click Continue

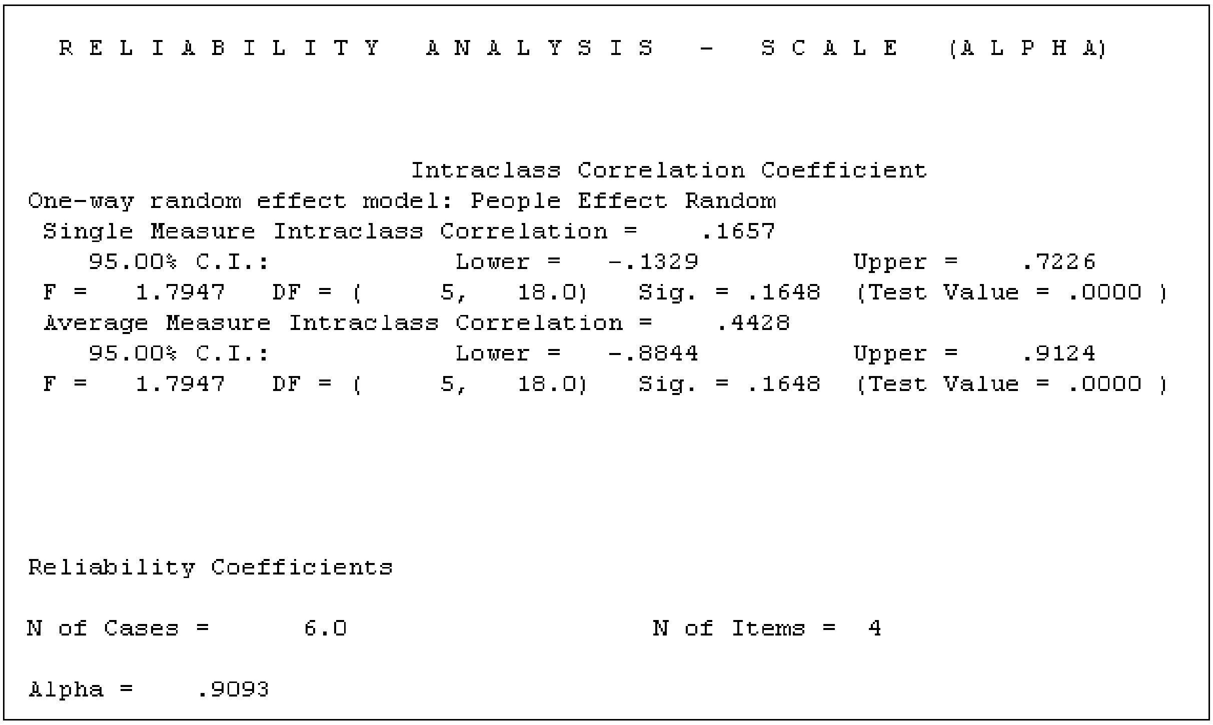


Items:
judge1
judge2
judge3
judge4

Model: Alpha

□ List item labels

Then OK to produce the results ….

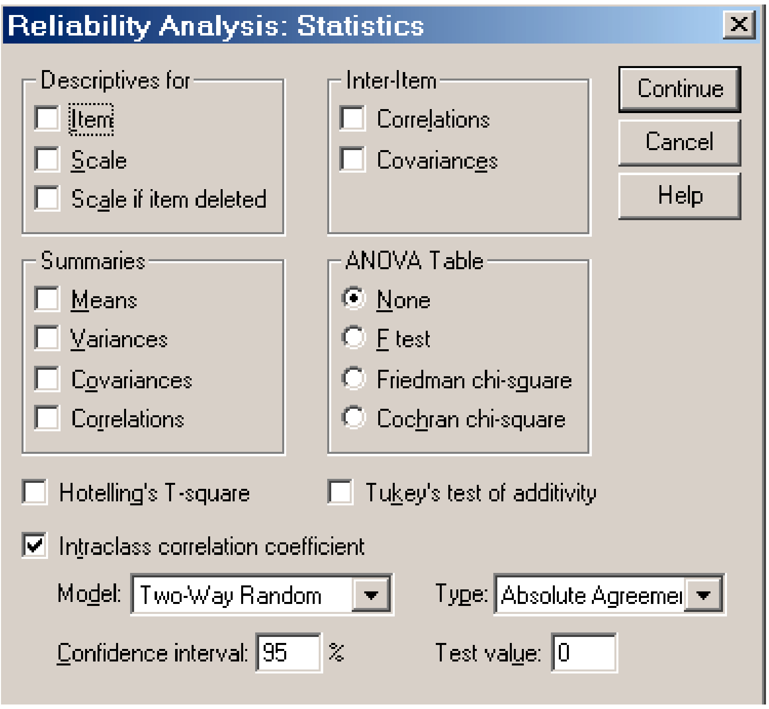


```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)



                         Intraclass Correlation Coefficient
One-way random effect model: People Effect Random
 Single Measure Intraclass Correlation =    .1657
    95.00% C.I.:            Lower =   -.1329          Upper =    .7226
 F =   1.7947   DF = (     5,   18.0)   Sig. = .1648  (Test Value = .0000 )
 Average Measure Intraclass Correlation =    .4428
    95.00% C.I.:            Lower =   -.8844          Upper =    .9124
 F =   1.7947   DF = (     5,   18.0)   Sig. = .1648  (Test Value = .0000 )





Reliability Coefficients

N of Cases =      6.0                   N of Items =  4

Alpha =    .9093
```

Thus Shrout and Fleiss Model 1 = SPSS One-Way Random model

Model 2: Every patient is rated by each rater. We assume the raters are randomly selected from some population of raters (a two-way random effects model). In essence, each rater rates all patients on all variables. This is the default model that covers most rating situations. For example, for a study in which we rate 10 patients on 5 variables, we would need at least 2 raters in order to assess interrater reliability. Each rater would make (10*5)=50 rating judgements.

Do everything as before until …Then select Intraclass correlation coefficient and **TwoWay Random** Model, and Type = Absolute Agreement



Reliability Analysis: Statistics

Descriptives for
☐ Item
☐ Scale
☐ Scale if item deleted

Inter-Item
☐ Correlations
☐ Covariances

Summaries
☐ Means
☐ Variances
☐ Covariances
☐ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

☐ Hotelling's T-square
☐ Tukey's test of additivity

☑ Intraclass correlation coefficient

Model: Two-Way Random
Type: Absolute Agreemer

Confidence interval: 95 %
Test value: 0

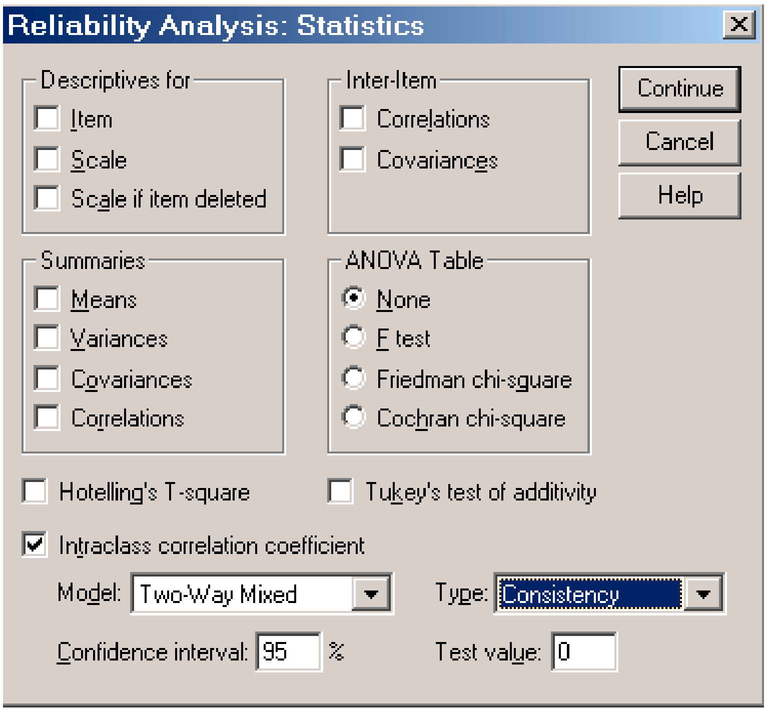
Continue and OK – for the results …

```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)



                        Intraclass Correlation Coefficient

Two-way Random Effect Model (Absolute Agreement Definition):
People and Measure Effect Random
 Single Measure Intraclass Correlation =     .2898*
    95.00% C.I.:             Lower =     .0188           Upper =     .7611
 F =   11.0272    DF = (      5,     15.0)    Sig. = .0001  (Test Value = .0000 )
 Average Measure Intraclass Correlation =     .6201
    95.00% C.I.:             Lower =     .0394           Upper =     .9286
 F =   11.0272    DF = (      5,     15.0)    Sig. = .0001  (Test Value = .0000 )
*: Notice that the same estimator is used whether the interaction effect
   is present or not.




Reliability Coefficients

N of Cases =       6.0                    N of Items =   4

Alpha =     .9093
```

Thus Shrout and Fleiss Model 2 = SPSS Two-Way Random model with Absolute Agreement

Model 3: Every patient is rated by each rater, **BUT**, in contrast to Model 2, **we assume the raters are THE population of raters** (a two-way, fixed-rater effects model). Each rater rates all patients on all variables. For example, for a study in which we rate 10 patients on 5 variables, we would select say 2 raters in order to assess interrater reliability. Each rater would make (10*5)=50 rating judgements. HOWEVER, it is assumed that these are the only two raters who will ever make ratings – **no generalizability is assumed to other raters.**

Do everything as before until …Then select Intraclass correlation coefficient and **TwoWay Mixed** Model, and Type = Consistency



Reliability Analysis: Statistics

Descriptives for
☐ Item
☐ Scale
☐ Scale if item deleted

Inter-Item
☐ Correlations
☐ Covariances

Summaries
☐ Means
☐ Variances
☐ Covariances
☐ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

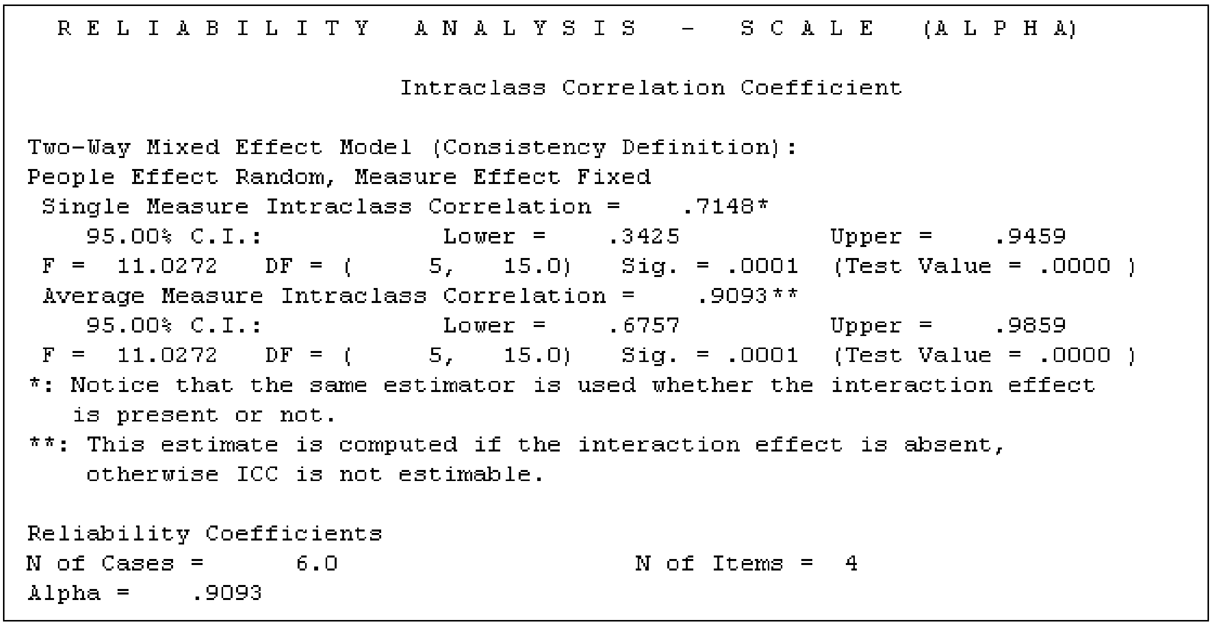
☐ Hotelling's T-square
☐ Tukey's test of additivity

☑ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Consistency

Confidence interval: 95 %
Test value: 0

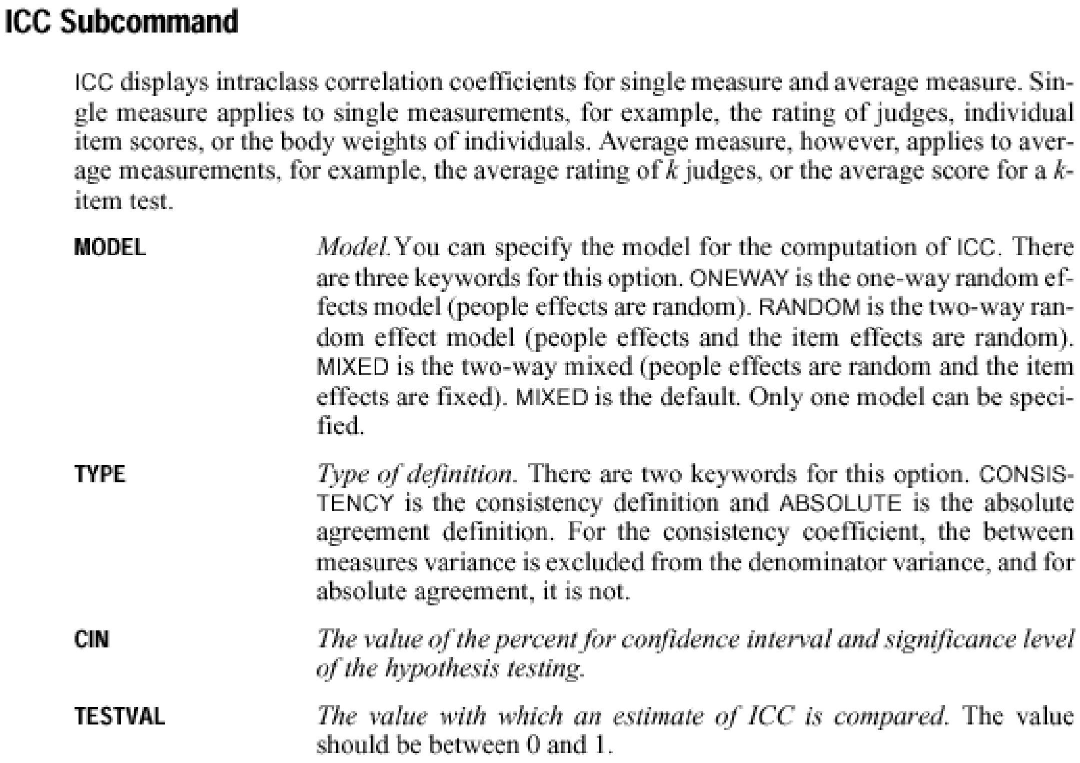
Continue and OK – for the results



```
  R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)

                       Intraclass Correlation Coefficient

Two-Way Mixed Effect Model (Consistency Definition):
People Effect Random, Measure Effect Fixed
 Single Measure Intraclass Correlation =     .7148*
    95.00% C.I.:             Lower =     .3425         Upper =     .9459
 F =   11.0272    DF = (      5,     15.0)    Sig. = .0001  (Test Value = .0000 )
 Average Measure Intraclass Correlation =     .9093**
    95.00% C.I.:             Lower =     .6757         Upper =     .9859
 F =   11.0272    DF = (      5,     15.0)    Sig. = .0001  (Test Value = .0000 )
*: Notice that the same estimator is used whether the interaction effect
   is present or not.
**: This estimate is computed if the interaction effect is absent,
    otherwise ICC is not estimable.

Reliability Coefficients
N of Cases =       6.0                    N of Items =  4
Alpha =     .9093
```

Thus Shrout and Fleiss Model 3 = SPSS Two-Way Mixed Model with Type = Consistency

From the SPSS 10.0 Base Manual – Reliability … ICC section …



ICC Subcommand

ICC displays intraclass correlation coefficients for single measure and average measure. Single measure applies to single measurements, for example, the rating of judges, individual item scores, or the body weights of individuals. Average measure, however, applies to average measurements, for example, the average rating of *k* judges, or the average score for a *k*-item test.

MODEL *Model.* You can specify the model for the computation of ICC. There are three keywords for this option. ONEWAY is the one-way random effects model (people effects are random). RANDOM is the two-way random effect model (people effects and the item effects are random). MIXED is the two-way mixed (people effects are random and the item effects are fixed). MIXED is the default. Only one model can be specified.

TYPE *Type of definition.* There are two keywords for this option. CONSISTENCY is the consistency definition and ABSOLUTE is the absolute agreement definition. For the consistency coefficient, the between measures variance is excluded from the denominator variance, and for absolute agreement, it is not.

CIN *The value of the percent for confidence interval and significance level of the hypothesis testing.*

TESTVAL *The value with which an estimate of ICC is compared.* The value should be between 0 and 1.

"**People Effects**" in the SPSS dialogs equate to **Patients** in my dialog.
"**Item Effects**" in the SPSS dialogs equate to **Raters** in my dialog

*** Note*** ..The "**between measures**" variance referred to in the paragraph above on Type is the MS_r Between Raters component that appears in the denominator of Model 1 and Model 2 calculations – but not Model 3.

What levels of Interrater/Intraclass r are considered acceptable?

Fleiss (1981) and Cicchetti and Sparrow (1981) from the medical fraternity state:

< 0.40 = Poor
0.40 – 0.59 = fair
0.60 – 0.74 = good
> 0.74 = Excellent

However, given an alpha internal consistency coefficient of < 0.70 is considered unacceptable for applied psychometric reliability indices, and alpha is related to intraclass r, then I can only conclude that the medical fraternity are setting limits far too low.

Realistically, values above about 0.7-0.8 are acceptable for applied tests. Below this value, and we have real problems using rating data. **Remember, the unconditional standard error of measurement for a rating scale is conventionally given by:**

$$SEM_x = s_T \cdot \sqrt{(1 - r_{xx})}$$

where

SEM_x = the standard error of measurement for test score X

s_T = the standard deviation of the test scores (from a normative group)

r_{xx} = the reliability coefficient

Let’s take some real UK PCL-R data. If our rater reliability is say 0.45, with a test standard deviation of 7, (a maximum score of 40), and a mean score of 17, and an observed score of 25, we have a **SEM of 5.19,**
with a 95% confidence interval of our true score of between 10 and 30.

If we had an interrater **reliability of 0.80**, with all other factors the same,
then our **SEM is 3.13**,
with a 95% confidence interval of our true score of between 17 and 30.

If we had an interrater **reliability of 0.90**, with all other factors the same,
then our **SEM is 2.21**,
with a 95% confidence interval of our true score of between 20 and 29.

By the way, the true-score confidence intervals are asymmetric – as per Nunnally (1978). See the **TRUESCORE** program (available from http://www.liv.ac.uk/~pbarrett/programs.htm) For the application of confidence intervals in change-score analysis.

Key References

Cicchetti D.V., and Sparrow, S.S.(1981) Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency*, 86, 127-137.

Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions, 2nd. Edition. New York: Wiley.*

Orwin, R.G. (1994) Evaluating Coding Decisions. In H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis*. Russell Sage Foundation, pp. 150-151

Howell, D.C. (1997) *Statistical Methods for Psychology*, 4th Edition. Duxbury, pp.490-493

Shrout, P. E., Fleiss, J. L. (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 2, 420-428