Guidelines to calculation by the free software and interpretation of the measures of disagreement applied to reliability studies

Elisabeth Svensson, Anders Avdic.

Reliability expresses the extent to which repeated assessments yield the same result, which means the level of agreement in paired data. In *intra-rater reliability* studies paired data are obtained by test-retest assessments by the same rater, and in an *inter-rater reliability* study the level of agreement between two raters is evaluated.

The statistical method by Svensson allows for a comprehensive analysis of the reasons for an observed disagreement in paired data from rating scale assessments. The method makes it possible to identify and measure systematic disagreement, when present, separately from disagreement caused by individual variability in assessments. In reliability studies it is important to consider these sources of disagreement, as they have different impacts on the quality of scales and raters. **Systematic disagreement** is related to the group and can be reduced or taken into account when the reason for such a disagreement is identified. A high level of **additional individual variations** in paired assessments indicates that the question and/or the scale categories do not fit well to the rater(s) or that the assessments are sensitive to disturbing factors of the test situation.

These guidelines refer to a worked example regarding evaluation of agreement and disagreement between two raters scoring the performance of 50 individuals by a five-point scale. The raters are denoted X and Y, and the categories A<B<C<D<E.
The figures and results come from the software.

a. The **frequency distribution** of the pairs of assessments is shown in the contingency table.

b. The diagonal of identical categories is marked. Ten out of 50 individuals were identically scored by the two raters, hence the **percentage agreement, PA**, is 20%, (see Results table).

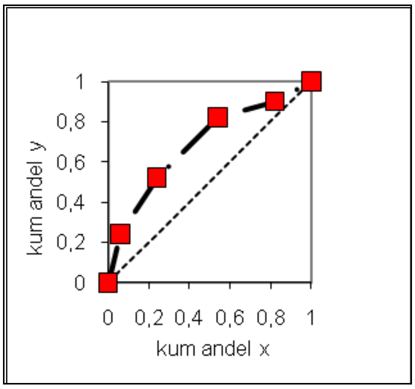
Contingency table		**X**						total
		A	**B**	**C**	**D**	**E**	**F**	
Y	**F**							0
	E				1	4		5
	D			1	1	2		4
	C		2	2	8	3		15
	B	1	1	9	3			14
	A	2	6	3	1			12
total		3	9	15	14	9	0	50

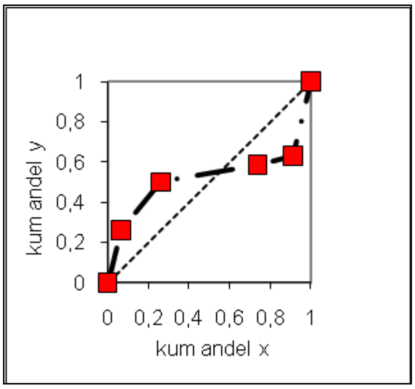
Figure 1. The frequency distribution of pairs of assessments by a five-point scale (A-E) of 50 individuals made by two raters (inter-rater agreement study)

WHY DO THE RATERS DISAGREE?

c. The marginal distributions show the frequency distributions of assessments on the five categories from rater X and rater Y, respectively. For example, according to rater X three individuals were scored category A, while rater Y scored 12 individuals in category A. Different marginal distributions are sign of **systematic disagreement** between the raters. In this example it seems that rater Y more frequently used lower categories.

The type of systematic disagreement can be illustrated by a **Q-Q-curve** (also called Relative Operating Characteristic, ROC curve), which is constructed by plotting the pairs of cumulated proportions against each other.
A concave or convex curve is a sign of systematic disagreement in position, and when the curve is S-shaped, the raters concentrate their assessments differently on the scale categories.





d. Two measures of systematic disagreement can be used**; the relative position, RP**, and the **relative concentration, RC.**

The RP expresses the extent to which the marginal distribution of rater Y is shifted towards higher categories than the marginal of rater, that is (X<Y), rather than the opposite (Y<X). A theoretical description of RP is the difference between the probabilities Prob(X<Y)-Prob(Y<X). Hence a positive RP value indicates that Rater Y has systematically more frequently used higher categories than X. Possible values of RP range from (-1) to 1.

The RC expresses the extent to which the marginal distribution of Y is more concentrated to central scale categories than the marginal of X, in short that (X<Y<X), rather than the opposite, (Y<X<Y). A theoretical description of RC is the difference between the probabilities Prob(X<Y<X)-Prob(Y<X<Y). Possible values of RC range from (-1) to 1, and a positive RC indicates that the assessments Y are more concentrated than X.

The shape of the Q-Q-curve and the RP-value (see Results table) confirm the presence of systematic disagreement in how the raters define the scale categories. RP = -0.38, and the negative RP means that rater X more likely used higher categories than did rater Y. Furthermore, in this data set it was 38 percentage units more likely that the individuals were scored higher by X than by Y rather than the opposite.
The 95 % confidence interval from -0.50 to -0.25, does not cover zero value of RP, which means a statistically significant disagreement is how the two raters use the scale categories (inter-rater bias).

In this example, the RC value is small and the 95% confidence interval covers the zero value, so the main reason for systematic disagreement is the systematic disagreement in position identified by the RP value.

Result table.

The measures of agreement/disagreement		Standard error (SE)	95% confidence interval	
PA	20%			
RP	-0,3768	0,0638	-0,5019	-0,2517
RC	-0,108	0,126	-0,36	0,14
RV	0,0648	0,0300	0,0060	0,1236
D	0,0763			

Now we know that one reason for disagreement is the systematic bias in how the raters interpret the scale categories. But is this bias the only explanation?

e. **The rank-transformable pattern (RTP)** of agreement is provided, when desired. The RTP shows the paired distribution that is expected is the case of systematic disagreement only, since it is completely defined by the two sets of marginal distributions (see Figure 2) The RTP confirms that rater X will systematically score one category higher than rater Y.

The rank-transformable pattern RTP		X						total
		A	B	C	D	E	F	
Y	F							0
	E					5		5
	D					4		4
	C			1	14			15
	B			14				14
	A	3	9					12
	total	3	9	15	14	9	0	50

Figure 2: The rank-transformable pattern defined by the marginal distributions of Figure 1.

f. It is obvious that the observed distribution of pairs differs from the rank-transformable pattern, RTP. This dispersion of pairs indicates additional presence of **occasional, individual-based disagreement.**

A typical sign of occasional individual disagreement is the disagreement in order between pairs of data. For example, in Figure 1 two individuals were scored the higher category E by rater X and D by rater Y, these pairs appear in the cell (E,D), these two observations are disordered the pair in the cell (D,E), since rater X scored lower than Y. The proportion disordered pairs out of all possible combination of pairs defines **the measure of disorder, D.** In the rank-transformable pattern all pairs agree completely in ordering and D = 0.
The calculations of the paired distribution of our example shows that D = 0.07, which means that 7% of all possible pairs are disordered.

g. The measure of disorder only counts the number of disordered pairs irrespective of the level of dispersion from the rank-transformable pattern. **The relative rank variance, RV**, is a rank-based measure of the observed individual variability and is defined by the sum squares of rank differences when the **ranks are tied to the pairs of observations** in the cells, so called augmented ranks.
In this case RV = 0.06
Like all measures of variance, the RV is hard to interpret, but RV<0.1 would in general be regarded as negligible.

Conclusion.
In the present example, the raters agreed in 20% of the assessments, and the disagreement were mainly caused by a systematic disagreement in how the raters interpret the scale categories. Rater X more likely scored one category higher than did rater Y (RP, -0.38; 95% CI -0.50 to -0.25) but additional occasional variations were found (RV, 0.06; 95% CI, 0.01 to 0.12).
This means, that the inter-rater reliability could be substantially improved by informing the raters about this bias and/or by training. The individual variability was negligible.

Read more:

Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. Statistics in Medicine 1994;13:2437-53

Svensson E, Starmark J-E, Ekholm S, von Essen C, Johansson A. Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. Neurological Research 1996;18: 487-94

Svensson E. A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical Journal 1997;39:643-57

Jonsson I, Gummeson L, Conner M, Svensson E. Assessing food choice: reliability and predictive validity of a method using food photographs in stacking boxes. Appetite 1998;30:25-37

Svensson, E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. Journal of Epidemiology and Biostatistics, 1998; 3 (4):403-409

Berntson L, Svensson E. Pain assessment in children with JCA; a matter of scale and rater. Acta Paediatrica 2001;90:1131-6

Svensson MH, Svensson E, Lasson A, Hellström M. Patient Acceptance of CT Colonography and Conventional Colonoscopy: Prospective Comparative Study in Patients with or Suspected of Having Colorectal Disease. Radiology 2002; 222:337-345

Svensson E. Statistical methods for repeated qualitative assessments on scales. Int J Audiol. 2003; 42 Suppl 1:13-22.

Allvin R, Ehnfors M, Rawal N, Svensson E, Idvall, E. Development of a questionnaire to measure patient-reported postoperative recovery: content validity and intra-patient reliability. Journal of Evaluation in Clinical Practice 2009;15:411-19