Agreement: I. Continuous measures

In many studies where some entity [e.g. a patient's symptoms; the part of a neuroimage corresponding to a brain region; some feature of an electroencephalogram (EEG)] is measured or rated on some property, the rating involves a degree of variability (often due to subjectivity), so that an important question is 'how reliable or trustworthy is the rating?'.

To answer the question researchers typically obtain two or more ratings of the entity and then, across a number of entities, examine the degree to which the different raters agree with each other. These ratings are typically independent, but not always, such as when a patient makes repeated ratings of their self in order to assess test-retest reliability. In the development of measurement instruments, such as a rating scale for the symptoms of a disorder, this question of *inter-rater* (or *inter-observer*) *agreement* is often expressed as one concerning *inter-rater reliability* and *test-retest reliability*.

The area of agreement (and reliability) uses differing definitions of the concepts and consequently different statistics can be calculated. A fundamental distinction can be seen by considering a group of students marked by two teachers. First, the teachers can agree by putting the students in the same order. Second, they can further agree by giving them exactly the same marks. If the focus of agreement is on the same ordering or ranking this is called *consistency* and if it is on the same values it is called *absolute agreement*. In some cases absolute agreement is important – consider pathologists rating biopsies where different ratings can lead to quite different treatment paths. In other cases consistency or reasonable similarity will suffice – consider two research nurses measuring blood pressure where exactly the same systolic/diastolic values are neither likely nor required.

Another distinction can be seen if we consider two columns of ratings that have been made on twins, that is the entities are twin pairs and we are interested in agreement between twins. With monozygotic twins it is arbitrary which twin is in the first column and which in the second. A similar situation arises when each entity is rated by, say, three raters, but the raters vary from entity to entity (e.g. three nurses on the acute ward rate each patient, but each nurse could be one of a dozen or more depending on who is on duty). In both cases we cannot distinguish between columns. On the other hand, if we restrict the data to dizygotic male–female pairs, then we can distinguish columns, since one column could be males and the other females. Similarly, on the acute ward the distinguishable raters might be the ward director, senior registrar and senior nurse.

The statistic used to measure agreement depends on the nature of the rating. Statistics for ratings that are clearly categorical (e.g. a symptom rated as *present* or *absent*; or as *worse, unchanged* or *improved*) will be looked at in a subsequent article. Here, we will look at ratings that are clearly continuous, such as functioning rated 0–100. Ratings using a limited number of categories (e.g. severity rated 1–5) are ambiguous and will sometimes be treated by researchers as categorical and sometimes as continuous. When a number of ratings (e.g. of symptoms) are added up to make a scale, then the resulting total score usually will be continuous, whereas the individual items usually are categorical, so that analysis of agreement for the scale will differ from that for the individual items.

The statistic usually used with continuous data is a correlation. The Pearson correlation has been traditionally used when reporting inter-rater or intra-rater reliability, though a researcher interested only in comparing rankings could use a Spearman rank correlation. Because neither the Pearson nor the Spearman are influenced by mean differences they only tell us about consistency.

An increasingly preferred option is the intra-class correlation [ICC; see Ref. (1) for a simpler presentation of details and (2) for a more complex discussion]. The ICC can be calculated using analysis of variance (ANOVA) and is applicable to any number of raters: this is the usual method. If there are only two raters, or raters are being compared in pairs, then an ICC can also be calculated using the *double entry* method, where a correlation is calculated after extending the data by adding on a copy, but with the columns reversed. This method is used by a number of researchers looking at twins or other pairings such as couples (so-called dyads). Pairs of raters is also the only way in which a single Pearson or Spearman correlation can be calculated.

Data where we cannot distinguish differences between the raters is typically analysed using a *one-way random effects* ANOVA model for the ICC (where random indicates that the entities are a random selection from all possible entities). Where we can incorporate differences between raters, a *two-way* ANOVA model is typically used. This model comes in two forms: one which focuses on consistency and one on absolute agreement.

Within the two-way model a further distinction can be made. If the raters are

https://doi.org/10.1111/j.1601-5215.2010.00445.x Published online by Cambridge University Press

Table 1. Different measures of agreement resulting from comparing Rater 0 with each of Raters 1–6 individually

		Raters compared against Rater 0					
	0	1	2	3	4	5	6
	40	40	45	60	45	52	55
	45	45	50	65	40	32	65
	50	50	55	70	50	62	45
	55	55	60	75	60	60	40
	60	60	65	80	55	48	50
Pearson		1.00	1.00	1.00	0.80	0.26	−0.58
Spearman		1.00	1.00	1.00	0.80	0.10	−0.60
Intra-class correlations (ICCs)							
2-Way consistency		1.00	1.00	1.00	0.80	0.24	−0.57
2-Way agreement		1.00	0.83	0.24	0.83	0.29	−0.81
1-Way		1.00	0.82	−0.23	0.84	0.34	−0.49
'Double entry'		1.00	0.78	−0.33	0.80	0.24	−0.57

random, in the sense of being just examples of the kind of raters (like whichever nurse is on duty), then we have a two-way *random effects* model (random entities, random raters). If the *specific* raters, however, are of particular interest (e.g. two different computer programs for making a diagnosis; comparison of a gold standard test and a new test) then it becomes a two-way *mixed effects* model (random entities, *fixed* raters). In practice, this last distinction is largely interpretational because the agreement statistics for the two models do not differ (2).

The entity being rated can be an individual or a number of individuals (as in a twin pair). The ratings for which we want to determine agreement can be a number of raters rating the entity, or the entity rated at different time or under different conditions, or the different parts of the entity (the two twins). The number of ratings has to be two or more and while we could have many ratings (columns of data) it is often impractical to have more than three or four.

In Table 1, we show some correlations for a small artificial dataset where we look at agreement between Rater 0 and each of the other raters. The raters are assumed to have rated patients on a 0–100 global assessment of functioning scale.

Rater 1 gave exactly the same ratings hence all the correlations are 1.0. Rater 2 made the same rankings, but rated each patient 5 points higher: as would be expected the first three correlations are 1, but the others pick up on the lack of total agreement. Rater 3 made the same rankings, but rated each patient 20 points higher: the clearly high lack of agreement is seen in the low ICCs – indeed two of them are now negative – however, the consistency-focused correlations remain at 1. Rater 4 shows some differences in rankings, but with scores on average closer than Raters 2 and 3: here the measures of consistency (the first three) turn out to be slightly lower than those of agreement, and while this might seem counter-intuitive, it emphasizes how we can differ in rankings yet still be relatively close in our absolute values. Rater 5 shows reasonable disagreement which is seen in the low values; we can also see differences between the various measures that are perhaps unexpected. Finally, Rater 6 shows strong disagreement, mainly by rating in the opposite direction.

Researchers are free to use whichever measure of agreement they prefer, provided they understand what agreement means under each measure and they appreciate how to interpret the resulting statistics.

Dusan Hadzi-Pavlovic[1,2]

[1] School of Psychiatry, University of New South Wales,
Kensington, NSW, Australia; and
[2] Black Dog Institute, Prince of Wales Hospital, Randwick,
NSW, Australia

Dusan Hadzi-Pavlovic,
Black Dog Institute Building,
Prince of Wales Hospital,
Hospital Road,
Randwick,
NSW 2031, Australia.
Tel: +61 2 9382 3716;
Fax: +61 2 9382 3712;
E-mail: d.hadzi-pavlovic@unsw.edu.au

Acta Neuropsychiatrica 2010: 22: 93–94
© 2010 John Wiley & Sons A/S
DOI: 10.1111/j.1601-5215.2010.00445.x

References

1. Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychol Bull 1979;**86**:420–428.
2. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;**1**:30–46.