

# Assessing Agreement: The Confusion Matrix

John Wills Lloyd

It is often important to assess whether one can trust the data in a research project. Were the scores from a test recorded accurately? Did two observers' records reflect the same observations. Did two people extract the same data when they coded studies for a literature review?

A valuable step in determining the trustworthiness of one's data is to assess inter-coder agreement, and a first step in doing so is to create a confusion matrix. A confusion matrix allows the researcher to assess the extent to which two scorers get confused—or are not confused!—about the definition of a given code in the scoring system. Here is an illustration of a simple confusion matrix.

		Scorer 2	
		“yes”	“no”
Scorer 1	“yes”	a	b
	“no”	c	d

To use this simple matrix, one cross-classifies each individual decision made by two independent scorers (observers, raters, scorers, judges, etc.). Imagine putting hash marks in the cells:

- ★ If both Scorers report that the event occurred, we count one in cell 'a.'
- ★ If Scorer 1 reports that the event occurred, but Scorer 2 reports that the event did not occur, we count one in cell 'b.'
- ★ If Scorer 1 reports that the event did not occur, but Scorer 2 reports that the event did occur, we count one in cell 'c.'
- ★ If both Scorers report that the event did not occur, we count one in cell 'd.'

Based on the data in these cells, one can calculate the agreement between the

*A useful aside:* House (1980) noted that by examining the off-quadrants ('b' or 'c') in a confusion matrix, one can ascertain whether one of the scorers is reporting biased data. If, for example, disproportionately more of the disagreements between the scorers are in the 'b' cell, then it must be that Scorer 1 is using a more lenient or generous definition to score the event than Scorer 2 is using.

scorers. The level of agreement is often used in assessing inter-observer agreement in observational research, inter-coder agreement when two or more individuals score studies in literature reviews, and other similar situations in scientific endeavors.

One can enter the counts from the matrix into formulae to assess the agreement between scorers. There are very basic alternatives (percentage agreement) that can be misleading and others that are more helpful (Hartman, 1977). To download a spread sheet that permits one to perform the calculations go to <http://bit.ly/agreecalculator>.

Of course, one can may need confusion matrices for decisions about three, four, or many more codes. Such extensions are not trivial but still not too hard. The second table illustrates the case when there are three possible codes on which the scorers might agree or disagree. (See, also, Bakeman and Gottman, 1997.)

		Scorer 2		
		“Code 1”	“Code 2”	“Code 3”
Scorer 1	“Code 1”	a	b	c
	“Code 2”	d	e	f
	“Code 3”	g	h	i

To use this more complex matrix, one still simply cross-classifies each individual decision made by the two independent scorers. The rules, however, are more extensive.

- ★ If both Scorers classified the event as Code 1, we count one in cell ‘a.’
- ★ If Scorer 1 classified the event as Code 1, but Scorer 2 classified the event as Code 2, we count one in cell ‘b.’
- ★ If Scorer 1 classified the event as Code 1, but Scorer 2 classified the event as Code 3, we count one in cell ‘c.’
- ★ If Scorer 1 classified the event as Code 2, but Scorer 2 classified the event as Code 1, we count one in cell ‘d.’
- ★ If both Scorers classified the event as Code 2, we count one in cell ‘e.’
- ★ If Scorer 1 classified the event as Code 1, but Scorer 2 classified the event as Code 3, we count one in cell ‘f.’

- ★ If Scorer 1 classified the event as Code 3, but Scorer 2 classified the event as Code 1, we count one in cell 'g.'
- ★ If Scorer 1 classified the event as Code 3, but Scorer 2 classified the event as Code 2, we count one in cell 'h.'
- ★ If both Scorers classified the event as Code 3, we count one in cell 'i.'

As should be obvious, in the case of this 3-x-3 matrix, the cells where the two scorers agree on the codes are 'a,' 'e,' and 'i.' The calculations of Kappa and Phi are more complex when one has multiple codes, especially when there are more than appear in this simple example. It helps to have computer programs, and both SAS and SPSS include procedures for performing such calculations.

If scorers are confused about codes, that is, if the levels of agreement are low, than the researcher responsible for the coding system should revise the coding system before proceeding with the research project. An untrustworthy measure is almost certain to lead to an untrustworthy study.

### References

- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.
- Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*, 103-116.
- House, A. E. (1980). Detecting bias in observational data. *Behavioral Assessment, 2*, 29-31.