Animal Behaviour 78 (2009) 1487–1491

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour

Allison B. Kaufman [a,*], Robert Rosenthal [b,1]

[a] Department of Neuroscience, University of California, Riverside
[b] Department of Psychology, University of California, Riverside

ARTICLE INFO

Article history:
Received 9 January 2009
Initial acceptance 2 April 2009
Final acceptance 3 September 2009
Available online 12 October 2009
MS. number: AS-09-00018R2

Keywords:
behavioural observation
Cohen's kappa
focused kappa
interobserver reliability
observational methodology
omnibus kappa
percentage agreement
reliability

Interobserver (or inter-rater) reliability is a vital part of all psychological studies that use an observational methodology to address questions of human behaviour. Concerns about reliability in these studies have long since left the arena of 'should we use an interobserver reliability statistic?' for debate on the particular type of statistic to be used, and academic careers have been built on this question. In stark contrast, however, it appears to be extremely rare to see interobserver reliability addressed at all in observational studies of animal behaviour. While we would never claim that this omission would or should be a basis for deeming a paper unacceptable, or disregarding its conclusions, we do feel that observational procedures are an integral part of the methodology of many studies, and that their inclusion in published papers should be commonplace. As an informal measure of the frequency with which interobserver reliability was addressed in papers involving the observation of animal behaviour, we surveyed articles recently published in the journal *Animal Behaviour*.

Our data came from volume 75 (3, 4) and volume 76 (1) of the journal, which were at the time the most recent issues available. We examined the first 100 articles (alphabetical by first author) that were methodologically relevant. Articles included in the survey used observational methodologies such as classification of behaviours, judgment of occurrence (or nonoccurrence) of behaviours, and the counting of instances of behaviour. Articles deemed not methodologically relevant and thus excluded from the analysis included studies using computer modelling techniques, studies in which results were strictly nominal (i.e. presence or absence of a physical object or the number of objects present), and studies that dealt with measurable quantitative variables such as weight, length or hormone levels. Studies such as these, of course, are also subject to error on the part of a single experimenter and are always improved by multiple, reliable experimenters; however, the problem is less pressing than it is in studies that deal with strictly behavioural observations.

Ninety-six of these 100 articles did not address interobserver reliability in their published text. Of these 96 articles, three mentioned some form of replication of the observations, seven

* Correspondence: A. B. Kaufman, Department of Neuroscience, LSP 2915, University of California, Riverside, Riverside, CA 92521, U.S.A.
E-mail address: allison.kaufman@email.ucr.edu (A.B. Kaufman).
[1] R. Rosenthal is at the Department of Psychology, 3111B Psychology Building, University of California, Riverside, Riverside, CA 92521, U.S.A. *E-mail:* robert.rosenthal@ucr.edu

doi:10.1016/j.anbehav.2009.09.014

specified using multiple observers but did not address reliability, 10 specified having used a single observer, and 76 articles made no mention of observational methodologies at all. Of the remaining four articles, two reported a percentage agreement statistic (Palagi 2008; Perry et al. 2008), one reported both a percentage agreement and a kappa statistic (Goossens et al. 2008), and one reported a kappa statistic based on two observers plus an additional observer that participated in 20% of the observations specifically for the purpose of establishing reliability (Riedel et al. 2008).

In and of themselves, the measurement of interobserver reliability and the discrepancy between the percentage agreement and the kappa statistic (to be addressed below) are in no way new concepts. These techniques are commonplace in statistics textbooks, with or without a focus on animal behaviour (see, for example, Martin & Bateson 1986; Bakeman & Gottman 1997; Rosenthal & Rosnow 2008). We provide a brief review of some important points for clarity's sake, as we believe the current treatment of these topics has become more a case of theory than of practise.

Methodologies and Techniques that Relate to Interobserver Reliability

Interobserver reliability

In general, there are two types of observer reliability: within observer reliability (i.e. consistency) and between observer reliability (i.e. interobserver). Studies that involve more than one observer should establish reliability between all observers to ensure that there is no specific bias on the part of any one person that might lead to bias in the data. Reliability can be affected by practise, experience, training, the rapidity of behaviour, the energy level of the observer or the clarity of a specific behaviour's definition. In addition, a single observer can be extremely consistent at measuring the wrong behaviour; however, there is no way to know whether this is the case without a valid comparison to another observer (Martin & Bateson 1986), and thus, the need for a comparison between observers to establish the accuracy of all observers.

In the event of an experiment in which only one observer is plausible, it is possible, and important, to assess reliability and estimate the likelihood of bias by briefly using a second observer to conduct interobserver reliability trials for a small, random, portion of the data.

Reliability within and between observers can be measured with a correlation between pairs of scores. It is, however, important to use multiple, preferably random, samples of behaviour in order to get an accurate assessment of reliability across all conditions of the study (Martin & Bateson 1986). In a comparison such as this, it seems intuitive that calculation of the number (or percentage) of times observers agree would be an appropriate measure. However, percentage agreement can be misleading, and several researchers in behavioural psychology have pointed out the inappropriateness of percentage agreement as a technique to measure inter-rater reliability (Suen & Lee 1985; Banerjee 1999; Lombard et al. 2002; Rosenthal 2005).

Percentage agreement statistic

Percentage agreement can often be a deceptive statistic: two observers may agree 98 out of 100 times that they have witnessed a particular behaviour, giving 98% agreement. However, the correlation between these observers' agreements can be a shocking $r = -0.01$. The observers may have obtained very good evidence of their ability to concur that this behaviour has occurred, however, we have no evidence of their ability to concur that a behaviour has not occurred. Conversely, if the two observers agree that a behaviour has occurred 49 out of 100 times and did not occur 49 out of a 100 times, their correlation is a much less surprising $r = 0.96$. This discord is apparent even when the data are less extreme, which is the more probable situation (Table 1).

Essentially, what we are seeing is due to a low level of variability in the judgments made by observers, which in turn makes percentage agreement a frequently misleading index of reliability. Unfortunately, percentage agreement is still commonly used despite its potential to mislead. In a review of articles published in the *Journal of Applied Behaviour Analysis* in which the researchers had used the percentage agreement statistic, data from 50–75% of these studies would have been deemed unreliable had a lenient kappa statistic (discussed below) been applied instead (Suen & Lee 1985).

The kappa statistic

One of the most common ways of measuring reliability between two observers without the problems inherent in percentage agreement is by using Cohen's kappa, which takes into account the chance agreement of two observers (Cohen 1960). It is thus a far more useful measure of interobserver reliability; kappa is defined as

$$\text{kappa}(k) = \frac{O - E}{N - E}$$

Where O is the number of times both observers agree, E is the number of times they would be expected to agree by chance, and N is the total number of observations. This calculation essentially removes the number of agreements due to chance from both the full set of observations and the subset in which both observers concurred. This leaves a chance-corrected proportion of observations.

The removal of chance agreement from the calculation of interobserver reliability ensures that the kappa statistic provides better evidence regarding any observational methodology, and will provide better evidence to support conclusions based on the behaviour of animals. Many animal behaviour studies use a single judgment on the occurrence of a behaviour, such as, whether a courtship display led to a successful mating attempt. Two observers may agree 90% of the time that successful mating occurred, but it would be misleading to draw any conclusions on unsuccessful mating based on the observers' failure to jointly define it.

Table 1
Examples of identical 98% and 50% agreement showing both negative and positive reliability

Agreement	Reliability				
98%	−0.01			*Observer 1*	
				Yes	No
		Observer 2	Yes	98	1
			No	1	0
98%	0.96			*Observer 1*	
				Yes	No
		Observer 2	Yes	49	1
			No	1	49
50%	0.33			*Observer 1*	
				Yes	No
		Observer 2	Yes	25	50
			No	0	25
50%	−0.33			*Observer 1*	
				Yes	No
		Observer 2	Yes	50	25
			No	25	0