Animal Behaviour 90 (2014) 273–280

Contents lists available at ScienceDirect

Animal Behaviour

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

Commentary

Observer bias in animal behaviour research: can we believe what we score, if we score what we believe?

F. A. M. Tuyttens [a,b,*], S. de Graaf [a], J. L. T. Heerkens [a], L. Jacobs [a], E. Nalon [a,b], S. Ott [b,c], L. Stadig [a], E. Van Laer [a], B. Ampe [a]

[a] Animal Sciences Unit, Institute for Agricultural and Fisheries Research (ILVO), Melle, Belgium
[b] Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium
[c] Faculty of Bioscience Engineering, Katholieke Universiteit Leuven, Belgium

ARTICLE INFO

Article history:
Received 28 October 2013
Initial acceptance 19 November 2013
Final acceptance 14 January 2014
Available online 12 March 2014
MS. number: 13-00901

Keywords:
animal welfare
behaviour scoring
cognitive bias
confirmation bias
double-blind experiment
expectation bias
information bias
observer bias
panting score
qualitative behaviour assessment

Most observers in behaviour studies are aware of relevant information about the animals being observed. We investigated whether observer expectations influence subjective scoring methods during a class practicum. Veterinary students were trained in recording negative and positive interactions between pigs, in scoring the degree of panting in cattle and in applying qualitative behaviour assessment (QBA) using a fixed set of terms for assessing hens' behaviour. The students applied these methods in three trials in which they were shown duplicated video recordings of the same animals: the original and a slightly modified version (to prevent recognition at second viewing). When scoring the duplicated recordings they were told either correct or false information about the conditions in which the animals had been filmed. The false information reflected plausible study scenarios in ethology and was used to create expectations about the outcome. As in reality the students scored the identical behaviour twice, the difference in the scores for the original and modified recordings reflects expectation bias due to providing different contextual information. In all trials there was evidence of expectation bias: students scored the ratio of positive to negative interactions higher when told that the observed pigs had been selected for high social breeding value, they scored cattle panting higher when told that the ambient temperature was 5 °C higher than in reality, and in the QBA they indicated more positive and fewer negative emotions when told that the hens were from an organic instead of a conventional farm. The magnitude of the bias in the QBA trial was related to the opinion of the students about hen welfare in organic versus conventional farms. Although veterinary students may not be representative of practising ethologists, these findings do indicate that observer bias could influence subjective scores of animal behaviour and welfare.

Scientific research that relies on observation and interpretation by the investigator has long been confronted with a well-known and fundamental problem: humans cannot be assumed always to process information objectively and accurately. Natural selection has shaped the human sensory processing system to promote behaviour that enhances the spread of our genes, not necessarily to provide us with a complete and correct picture of reality. There is ample evidence that human perception can be selective and biased, and that a healthy human brain often makes incorrect associations and deductions (Braeckman & Boudry, 2011). For example, psychologists have long recognized that people are prone to expectation bias, which refers to the psychological sway towards one opinion versus another as a result of possessing information extraneous to the task at hand (Page, Taylor, & Blenking, 2012). Information that confirms one's beliefs or hypotheses is often favoured (Nickerson, 1998; Wason, 1960). Investigators and research staff are also susceptible to these pitfalls of the human brain. Often, they carry out experiments while they are predisposed by strong expectations about the outcome and deep-rooted assumptions about what is and what is not possible. These expectations may lead to conscious or unconscious biases in observation and recording of data.

The risk of these types of observer bias can be reduced by ensuring that the person collecting the data is unaware of which treatment each subject has received until after the experiment. Such blind trials are widely considered as the best study design to minimize observer bias and are often required to attain regularity approval for new drugs, dietary supplements and medical

* Correspondence: F. Tuyttens, Animal Sciences Unit, Institute for Agricultural and Fisheries Research (ILVO), Scheldeweg 68, 9090 Melle, Belgium.
E-mail address: frank.tuyttens@ilvo.vlaanderen.be (F. A. M. Tuyttens).

http://dx.doi.org/10.1016/j.anbehav.2014.02.007
0003-3472/

devices (Kaptchuk, 2001; Miller & Stewart, 2011). Meta-analyses have convincingly shown that randomized trials have substantially larger treatment effects if the assessor is not blinded (Hróbjartsson et al., 2012, 2013; Schulz, Chalmers, Hayes, & Altman, 1995). Nevertheless, the use of nonblinded assessors in nonpharmacological trials remains common (Hróbjartsson et al., 2013; Schulz et al., 1995), particularly in animal behaviour studies. Burghardt et al. (2012) reviewed several hundred articles published in five leading animal behaviour journals during the last five decades. They found that, despite the numerous and widely used texts on research methods in animal behaviour that advocate researchers should minimize potential observer bias in their studies, only 6.3% of the empirical research articles reported that at least one component of the research was conducted blind. This percentage was much higher in two more human-focused comparison journals that publish research based on similar behavioural observations and coding strategies (25% for *Behavioral Neuroscience* and 47.5% for *Infancy*). We checked the 2012 volumes of *Animal Behaviour* and *Applied Animal Behaviour Science*; only 15.3% (37/242) and 9.9% (13/131) of the papers for which we judged it relevant reported that they were conducted blind.

Marsh and Hanlon (2004) remarked that while the potential for observer biases in animal behaviour research has often been discussed, there have been few quantitative analyses of the kinds of biases that may affect behavioural data. We cannot find any scientific reason to justify the limited attention being paid to minimizing observer bias in animal behaviour science. Observer bias is particularly likely when the investigator has strong preconceptions or a vested interest in the outcome, when the underlying data are ambiguous and when the scoring method is subjective (Hoyt & Kerns, 1999; Page et al., 2012; Risinger, Saks, Thompson, & Rosenthal, 2002; Rosenthal, 1966, Schulz, Chalmers, & Altman, 2002). In our opinion, all three predisposing factors are commonly present in animal behaviour research.

In this study, we investigated the potential of observer bias in animal behaviour studies. Veterinary medicine students unknowingly participated in an experiment during a class practicum. The students were first given a brief demonstration of several subjective (in the sense that they rely on an individual's perception and judgement, and can therefore be influenced by experience or personal views, cf. Meagher, 2009) animal-based scoring methods commonly used in ethology. They then received a short training session on using these methods. Last, they applied the methods to score video clips of farm animals that they were led to believe had been subjected to different conditions. In reality, however, they scored the same video clips twice, with the second clip being slightly modified to trick the students into thinking that the videos were of different animals subjected to the mock conditions. The false information was specifically chosen to create expectations among the students about the outcome of the observations. The research objective was to calculate the magnitude of observer bias as the difference between the first and second scoring of the video clips. In addition to this research objective, the practicum was also designed to meet two educational goals: to learn and practise a limited but diverse set of scoring methods used in ethological research, and to raise the students' awareness about observer bias by allowing them to experience how prior expectations affected their scoring. Goldstein, Hopkins, and Strube (1994) argued that a classroom demonstration in which the students personally experience the powerful effects of previous expectations on perception should improve their learning and memory about observer bias. We do not include that aspect of this study in the current paper and report on the research objective only.

METHODS

Subjects and Overall Experimental Set-up

The trials were conducted using third-year veterinary medicine students at Ghent University during the practicum of the Ethology, Ethics and Animal Welfare course. The students ($N = 157$) were informed that the aim of the practicum was to learn some techniques and methods commonly used in ethological research: (trial 1) identifying negative and positive social interactions in pigs; (trial 2) scoring panting as an indicator of heat stress in cattle; (trial 3) qualitative behaviour assessment (QBA) of laying hens. For logistic reasons, the training and trials were based on photographs and video recordings rather than live observations. These video recordings were obtained from recent or ongoing research projects approved by the ILVO ethical committee for experiments on animals. All methods concerned animal-based measures that were subjective in that they required some interpretation and judgement by the students when giving scores or values to their observations of the animals. The students were not informed about the research aim of the practicum, namely to investigate the extent to which they were affected by misleading background information when applying the three scoring methods they had learned.

After the students had been informed about the content and organization of the practicum, they were randomly split into two groups of approximately equal size. For the training and the three trials group A ($N = 80$) remained in the auditorium, whereas group B ($N = 77$) moved to another nearby auditorium. Each trial was chaired by a different designated pair of researchers and lasted 30–40 min followed by a 10 min break. The order of the sessions differed to enable the designated researchers to repeat their trial in both places with both groups. After the scoring sessions, the two groups were reunited for a plenary session during which they were informed about the hidden research aim of the practicum and the reason why we investigated this subject (i.e. to raise awareness about the pitfalls of their own senses and brain when applying subjective animal-based measures for assessing animal behaviour and welfare).

Trial 1: Negative and Positive Social Interactions

The students were taught how to recognize negative and positive social interactions in fattening pigs (using video recordings of various interactions) and were briefly informed about the relevance of these interactions for animal welfare and farm management. They were taught an ethogram that included four types of negative interactions (head butt, belly nosing, ear or tail biting, biting other body parts) and three types of positive interactions (play, sniffing/nosing the body of a pen mate, nose-to-nose contact). During a video-based training session, the students then practised counting social interactions, differentiating between positive and negative ones.

Subsequently, for the trial proper, the students were instructed to tally each negative and positive social interaction in a pen of six fattening pigs during two 5 min video clips. Both video clips were the same, but the second clip was shown in a mirror image, its brightness was slightly adjusted and fictional pen numbers and dates were shown. The goal of showing the same clip twice (once in the original version and once in the slightly modified version) was to mislead the students into believing that it concerned two different groups of pigs from an experiment on social breeding value and housed in an opposite pen of the same pig stable. The theory of social breeding value, in which pigs with a high social breeding value have a positive effect on the growth of their pen mates (Bergsma, Kanis, Knol, & Bijma, 2008), was explained. The