Applied Animal Behaviour Science 119 (2009) 1–14

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

Review

Observer ratings: Validity and value as a tool for animal welfare research

Rebecca K. Meagher *

Animal & Poultry Science Department, University of Guelph, 50 Stone Road East, Building #70, Guelph, Ontario, Canada N1G 2W1

ARTICLE INFO

Article history:
Accepted 16 February 2009
Available online 21 March 2009

Keywords:
Animal welfare
Scoring methods
Non-invasive
Subjective assessment
Behavioural assessment

ABSTRACT

Ratings by human observers have long been used by animal scientists and veterinarians to assess certain physical traits (e.g. body fat), and can also be applied to the assessment of behaviour and a variety of welfare-relevant variables (e.g. pain responsiveness, alopecia/barbering). Observer ratings offer a myriad of advantages, not just practical (e.g. money-saving) but also scientific: they can be used to integrate multimodal information across time and situations, and for constructs that are otherwise very difficult to assess (e.g. nest quality). Because observer ratings involve subjective judgements, some researchers may question whether they can be trusted to reflect reality in an unbiased manner. In this paper, I present evidence from a range of zoo, laboratory and farm animal studies demonstrating that observer ratings can be both reliable and valid. They have been shown to predict important biological phenomena such as reproductive success in rhinoceroses and cheetahs. Biases are indeed a risk, particularly when the ratings could reflect on the observer's own care of the animals or on their institution; however, this risk can be minimized through careful experimental design, including blinding and careful phrasing of the questions the observers need to answer. I review the steps involved in validating an observer rating scheme, and also discuss both study design issues (e.g. selecting terms to be rated and appropriate observers) and the statistical issues some schemes may raise (e.g. ordinal data are not truly normal).

Contents

* Tel.: +1 519 824 4120x53557.
E-mail address: rmeagher@uoguelph.ca.

0168-1591/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.applanim.2009.02.026

1. Introduction

The idea of using ratings by observers in scientific studies is not new, but this method continues to be used much less frequently than are other methods in applied ethology. For example, traditional ethogram-based methods, which quantify specific behavioural elements in terms of frequencies and durations (Carlstead et al., 2000), have consistently been the focus of more papers, as shown in Fig. 1. Observer ratings are scores given for a variable, using units of measurement that are defined by the researchers, in contrast to standardized units such as those used to measure weight, or the counts used in the more traditional ethological methods. The units set by the researchers are often open to some degree of interpretation by the rater. The variable could be a behaviour pattern, a personality trait, a physical condition such as alopecia (Honess et al., 2005), or even a product of behaviour such as the quality of a nest built (e.g. Deacon, 2006). When multiple behaviours or attributes are being rated, they are compiled onto a single form on which each item receives a score. I will typically refer to both single- and multi-item measurement instruments as 'scales', since this is the most widely used term; however, it should be noted that some social scientists use the term 'scale' only when there are predicted relationships between the items included (i.e. scores are to be based on *patterns* of responses), while any other multi-item instrument that measures a single construct (i.e. the overall score is simply the sum of all items) is called an 'index' (Babbie and Benaquisto, 2002). Another term commonly used in psychology, if the ratings are given by someone other than the researcher, is a 'questionnaire'; I will use this term only when talking specifically about scales distributed by the researchers to be filled out by other observers.

In giving these ratings, the observer plays a more active interpretive role than would be required in traditional methods; he or she must consider and weight the relative importance of multiple pieces of information, sometimes gathered over a long span of time, to synthesize them into a single score (Block, 1977). The researcher may provide a set of terms on which to rate the animal. Alternatively, in the Free Choice Profiling (FCP) method (e.g. Wemelsfelder et al., 2001), the observer's role can also extend to choosing the vocabulary with which to describe what he or she observes (i.e. the terms on which the animals will then be rated); the risks of this method will be discussed below. Because of the active role played by the observer, data obtained in this way are sometimes referred to as 'subjective ratings', their being subjective in the sense that they rely on an individual's perception and judgement, and can therefore be influenced by experience or personal views.

In human psychology, subjective ratings and observations have long formed a standard component of research and clinical assessments. These include ratings by a variety of observers including clinicians, caregivers and peers, which complement or substitute for self-reports; for example, the use of staff ratings of psychiatric inpatient function dates back to the 1970s (Hersch et al., 1978). Observer ratings are of particular importance when working with people who cannot provide trustworthy self-reports (McCrae and Weiss, 2007), such as young children who have not yet mastered language, or patients whose disorders impair their cognitive function or ability to communicate. In animal science, observer ratings were first used for physical traits, sometimes using and validating scales that had been developed for producers (e.g. body fat: Russel et al., 1969; use in production: Jefferies, 1961). Scoring systems have also been developed for health-related variables such as lameness (e.g. Kestin et al., 1992; Flower and Weary, 2006) and pain (reviewed in Hawkins, 2002) that are relevant to veterinary research and welfare auditing schemes. These are regularly used informally, such as in daily checks of laboratory animals by animal care personnel. Stevenson-Hinde and Zunz (1978)

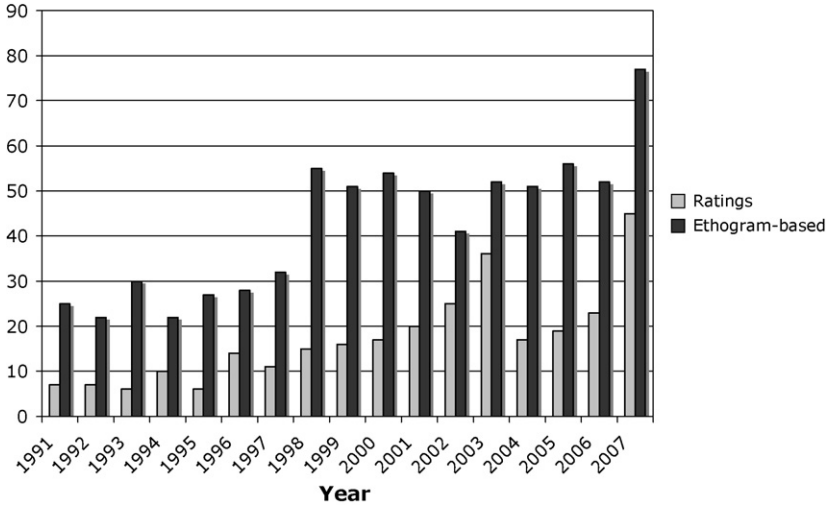

Fig. 1. Papers using observer ratings vs. ethogram-based methods. Journals included: *Animal Welfare*, *Applied Animal Behaviour Science*, *Veterinary Research*, *Laboratory Animals* and the *ILAR Journal*. Search terms were 'observer ratings' or "subjective assessment" or "scor*" for the first method, and "behav*" plus "freq*", "durat*" or "ethogram" for the second. Search conducted on Web of Science June 17, 2008.