

Caretaker ratings of animal behaviour

– a tool for animal welfare research

Rebecca Meagher

This paper was inspired by a visit and seminar by Dr. Kathy Carlstead, a researcher based at the Honolulu Zoo. As part of her presentation, Dr. Carlstead talked about how the knowledge of keepers could be use to identify individuals at risk of poor welfare. This approach appealed greatly to Becky, although in her paper, she is careful to assess its potential drawbacks too.

Introduction

Subjective ratings by observers, who could be scientists or animal handlers and caretakers, have been suggested as a tool for animal welfare research (Wemelsfelder, 1997). Indeed, this technique has already become common in the related area of animal personality (Jones & Gosling, 2005), and has parallels in the human literature. By obtaining ratings from owners or caretakers, scientists can take advantage of their intimate knowledge of the individuals (Carlstead, 1999a), as well as reaping other practical benefits. Some scientists would dismiss this method because it is described as subjective, and is thus believed to be too prone to biases and implicit, unjustified assumptions. It can be valuable only if, despite its heavy reliance on human perception and judgement, it produces reliable and valid data. A review of the literature demonstrates that this is possible. With careful questionnaire design and proper testing, caretaker ratings could provide a wealth of information that would otherwise not be easily available to scientists.

Subjectivity in science

The scientific method is centred on empirical observation. It is considered objective, meaning independent of personal opinions, and therefore providing an accurate representation of reality. In practice, however, all science depends on human perception and interpretation, to varying degrees. While it may be relatively simple for a physiologist to obtain a quantitative measure of the amount of a given compound in the body, animal behaviour and welfare researchers often deal with more holistic concepts than single, physical entities. Before any behaviour can be quantified, humans must define it, which involves some level of interpretation. Not only is subjectivity thus present in

the methods, but some research is also focused on subjective experiences such as “suffering”, which can be defined only from the perspective of the animal. Many scientists question the validity of all such research (e.g. Vanderwolf, 1998; reviewed in Fraser, 1999) because the inner experience of another is not open to empirical measurement. Ultimately, however, welfare science cannot avoid the topic of subjective experience, since concern about it constitutes the foundation of the field. It can be argued that similar attributions regarding inner experience are necessary in much of human psychology, and are less often examined critically in that context (Schilhab, 2002). There are scientifically accepted methods of drawing inferences regarding the subjective states of others when direct access is not possible.

What complicates the study of subjective experience in non-human animals (henceforth, referred to simply as “animals”) beyond that of human psychology is the concern with anthropomorphism. Anthropomorphism, the attribution of human characteristics to non-humans, is a term typically used as censure. Criticisms on these grounds have influenced the behavioural literature. For example, many animal researchers avoid the term “personality” purely for this reason, instead referring to “temperament” (discussed in Gosling, 2001) or even “behavioural types” (e.g. Gold & Maple, 1994) to describe concepts that would be called personality in the human literature. The view of anthropomorphism as intrinsically wrong, however, stems from philosophy rather than empirical evidence (Keeley, 2004). Accusations of anthropomorphism are often anthropocentric, based on unproven ideas about what makes humans unique (Heidegger, 1984, discussed in Tyler, 2003). While unthinkingly assuming that animals think or feel the same way people do is unscientific and can lead to erroneous conclusions, it is an equally grave fallacy to blindly deny the possibility of an attribute being shared by another species. As Panksepp (2003) argued, much evidence of genetic and neurobiological homology across species has been collected in recent decades. With adequate evidence to support the conclusion that an attribute or state is present in another species, it is logical to accept it rather than to risk overlooking important truths (Lehman, 1992). Backed by such arguments, fields such as cognitive ethology have embraced the concept of continuity between the minds of humans and other species, which has been implicitly accepted in psychology for decades (Keeley, 2004). Anthropomorphism, then, can be justifiably applied but only within the constraints of the scientific method.

From subjective concepts to subjective methods

As a result of the controversy surrounding these topics of investigation, nonetheless, animal behaviour and welfare researchers tend to be particularly concerned with upholding the standards of objectivity. Failure to do so carries the risk of the entire field not being seen as a legitimate science. It would be difficult to argue that ideas about subjective states have no scientific merit, since they generate testable hypotheses. It may be, however, that we will never be able to access these states directly, but only infer their existence. Standard physiological welfare measures such as corticosteroid levels are not sufficient for such inferences. They are influenced by many factors, making them difficult to interpret (e.g. Millspaugh et al., 2001; reviewed in Rushen, 1991), and not all factors leading to increased corticosteroids would be expected to have the same affective valence. Complementary analyses of behaviour are therefore necessary for any understanding of the animal's experience.

Wemelsfelder (1997) has argued that the study of subjective states calls for the use of qualitative methods of assessment, because these states are dynamic. She claims that they cannot be captured by examinations of individual movements or the condition of the animal at any given instant; rather, one must examine the "flow" of behaviour. Although there are, in fact, means of analysing sequences or overall patterns of behaviour quantitatively, these methods may miss subtle, idiosyncratic cues, some of which are picked up by qualitative observations (Carlstead et al., 2000). The use of integrative qualitative descriptors can also be more efficient than breaking behaviour down into measurable components. Wemelsfelder and colleagues use the term "qualitative assessment" for what many others would call "subjective assessment", because the method involves rating subjects on a set of such qualitative descriptors rather than quantifying the performance of specific behavioural elements. The terms can be provided or be generated by the observers in the initial phase of an experiment (e.g. Wemelsfelder et al., 2001). The risks of the latter method will be discussed below. Conveniently, the "qualitative assessment" label also avoids the use of the word "subjective", which might cause some readers to dismiss the work as inevitably unscientific. While it would be unscientific if the observations reported were based on the untested interpretations of one person or group, the method need not rest on individually variable perspectives and can be validated empirically. If, in addition to being more efficient, this method is valid, it may be a very appropriate tool for certain welfare studies.

In human psychology, subjective ratings and observations form a standard component of research and clinical diagnostic procedures. These include self-evaluations as well as ratings by a variety of observers including clinicians, caregivers and peers. Caregiver ratings are of particular importance when working with young children who have not yet mastered language, or with patients whose disorders impair their cognitive function or ability to communicate. One application of caregiver ratings is the assessment of the quality of life of medical patients. Quality of life, although often not clearly defined, is a measure of subjective well-being that incorporates both physical and mental health and, typically, satisfaction with one's perceived position in life (e.g. WHO: Saxena & Orley, 1997). It has been likened to the concept of welfare in animals (Fraser et al., 1997). Subjective ratings by observers are also relied on for diagnostic purposes, such as scoring pain (Hawkins, 2002) or lameness (Hewetson et al., 2006) in animals. The diagnostic accuracy of such lameness scores does not seem to be significantly increased by the use of a more objective scale (Flower & Weary, 2006).

The ubiquity of qualitative ratings is not the only reason to believe that they should be accepted. Block (1961, cited in Jones & Gosling, 2005) defends the use of ratings by observers, by demonstrating that combining ratings from multiple assessors largely eliminates individual idiosyncrasies, and thus minimizes the subjectivity of the measure. Additionally, each rating scale in psychology and medicine has been validated as it came into common usage. Jones and Gosling (2005) made the point that many other methods, by contrast, have never been tested for reliability, which means the repeatability of the measurement obtained; they are simply assumed to be reliable because they appear to be objective, but research in humans has demonstrated that this assumption is not always true. There is always the potential for errors or disagreements on the definition of the behaviour being measured. In reality, it is impossible to remove all subjectivity from the assessment of behaviour or temperament, since all methods depend on human perception. The optimal solution seems to be obtaining data from multiple observers, and taking steps to avoid any systematic biases in judgement. It should be possible to adapt psychologists' methods of constructing valid questionnaires for use in the study of animal behaviour and welfare.

Study Design

Writing questionnaires

There are many things to consider in composing and implementing the questionnaires used in this method. During the initial creation, one of the prime concerns is achieving a balance between simplicity and the use of multiple items related to one phenomenon. If the questionnaire becomes too complex or too time-consuming, the desired respondents will be less likely to participate or to give thoughtful, accurate responses (Carlstead et al., 2000; Martin, personal communication, 2006), possibly leading to invalid results. This is particularly true for ratings that are to be recorded on a daily basis rather than being completed on a single occasion. However, the use of multiple items can account for individual differences in the expression of a trait or phenomenon, and allow checks of internal reliability (see Wright & Feinstein, 1992 for a more detailed discussion). Although the former is less often necessary in the context of animal research than on diagnostic questionnaires, since most of the qualitative terms being rated are broad descriptors of a construct rather than specific manifestations, internal reliability might be a consideration when several distinct behavioural patterns are believed to share a common cause.

Clarity of the terminology is also essential. Ideally, multiple scientists should collaborate on the writing process, to ensure that the phrasing of the questions is unambiguous, and would be interpreted in a similar way by most people (e.g. Carlstead et al., 2000). If this is not achieved and people involved in the study are using a term in different ways, comparisons between ratings by different observers are impossible, and the researcher may draw incorrect conclusions regarding the underlying states or experiences of the animal. The careless choice of qualitative terms may allow unsubstantiated anthropomorphism to enter a study. The terms that come to mind most readily or are most easily understood by everyone involved may be those which are in the vernacular because they are meaningful in discussing human experience. However, terms used to describe a human state may not be appropriate for a superficially similar state in the species being studied. Herein lies the danger of allowing the untrained observers to generate their own terms and grouping them by similarity of meaning and use during the analysis phase. This process does not involve any testing of assumptions and may result in the use of terms which have not been validated in any way for the species of interest. The choice of terms should involve careful thought, taking into account current knowledge of the species.

Testing questionnaires

Once the questionnaire has been put together and the meanings of the terms agreed upon, it can be applied. However, before any conclusions can be drawn from the ratings, it must be validated. In order to do this, multiple observers must use it to rate a sample of animals that is representative of the population to which the results are intended to apply, and these ratings be subjected to statistical tests. There are several types of validity, described below, and although all types have usually not been systematically tested for any single questionnaire, an examination of the literature as a whole provides evidence that it is possible to obtain reliable, valid results with this method.

The first step in the validation process is typically testing reliability, both inter-observer and intra-observer. The former refers to agreement between multiple people independently rating the same individual, usually at approximately the same time. Several studies have confirmed that subjective ratings of animal behaviour and temperament can have reasonable inter-observer reliability (e.g. Carlstead, 1999a; Feaver, 1986; Wielebnowski, 1999). Intra-observer reliability, by contrast, is the agreement between ratings by the same individual on multiple occasions. Some authors restrict this term to comparisons of ratings of the same sample of behaviour from a video (e.g. Martin & Bateson, 1993), while others include comparisons of ratings from multiple observation periods, which could also be called test-retest reliability. Although fewer studies in this field have considered this type of reliability, those that have show high intra-observer agreement (Rousing and Wemelsfelder, 2006; Wemelsfelder et al., 2001). Carlstead and colleagues (1999a) were unable to test intra-observer reliability in keeper ratings of rhinoceros behaviour, because some keepers changed between observation periods; however, this study demonstrated test-retest reliability.

In some cases, imperfect reliability is not due to a flaw in the tool. An animal may interact differently with different people, leading to different ratings and, consequently, lower inter-observer reliability. This variability in behaviour should be captured in any overall assessments of the animal. It could be significant in welfare studies, since it is adaptive to alter responses depending on past experience with an individual or situation and thus, a lack of variability might indicate a problem. If there is poor agreement between observers on an item, it should be omitted from any analyses of relationships between the ratings and other measures, but would be worth further investigation. This would begin with having a third party rate the behaviour from videotapes, blinded to the identity of the observer present when each video was taken, to

determine whether the animal truly behaves differently in the presence of different observers. Follow-up studies could then analyse what attributes of the person or handling style elicit responses that are indicative of poor or good welfare.

Similarly, perfect test-retest reliability is not expected if the tests are spread over a long period of time. Animals are not static entities; their behaviour, even in a well-defined test situation, can be altered over time as a result of learning or may depend on physical or mental states. For studies of personality, this type of reliability should be high, since personality is defined as a set of individual characteristics that is consistent across time and situations (Gosling, 2001). By contrast, studies of animal welfare might examine changes in behaviour over time to determine the long-term effects of different management practices or other variables. Caretaker ratings are intended to integrate behaviour over a period of time and a variety of situations rather than reflect only a particular moment in time, and thus are expected to display some stability; however, if repeated years later or after a major intervention or other event in the life of the animal, some changes would be expected. Dependent on what is being measured, then, lack of reliability may or may not be a sufficient reason to discard data obtained from subjective ratings.

If there is significant concordance between ratings by independent observers, and these ratings are consistent across multiple observation periods and differ between individuals studied, they are likely to be measuring a real phenomenon. However, in order to be certain that it is actually the phenomenon of interest, tests of validity must also be conducted. Validity can be categorized into several broad types, such as content, criterion and construct (Cronbach & Meehl, 1955). In order for a questionnaire to have content validity, all items included must be relevant to its aim and the set of items must be sufficient for detecting all possible expressions of the phenomenon to be measured. This is typically a concern during the questionnaire design stage discussed above. Either criterion validity, the accuracy of the tool in predicting the score on a criterion measure, or construct validity, its ability to measure a “postulated attribute” such as a personality trait (Cronbach & Meehl, 1955), is tested later.

Criterion validity can be further subdivided into predictive and concurrent validity, depending on the timeline; if the criterion measure and that being validated are taken at roughly the same time, it is called concurrent, while predictive validity refers to the measure predicting the criterion measurement obtained some time later (Cronbach & Meehl, 1955). One example of establishing predictive validity is the study of future guide dogs by Serpell and Hsu (2001), who

used questionnaires completed by the volunteers who raised the puppies to predict which dogs would later be rejected as guide dogs during training.

In studies relevant to animal welfare, it is more common to investigate construct validity rather than criterion validity, because there is often no ‘gold standard’ criterion measure with which to compare the rating scale. As discussed above, animal welfare frequently deals with subjective constructs such as “suffering”, for which there is no perfect, comprehensive measure since all are indirect measures of the animal’s experience.

Construct validity is typically divided into convergent validity, which describes how well the measure correlates with others to which it is expected to correlate, and discriminant or divergent validity, which is the lack of correlation with measures to which it is not conceptually related (Jones & Gosling, 2005). The study by Serpell and Hsu (2001) discussed above demonstrated discriminant validity, as there was no correlation between rejection of a potential guide dog and ratings on behaviours that were unrelated to the reason for rejection. Hsu and Serpell (2003) also demonstrated discriminant as well as criterion validity of an owner questionnaire for assessing behaviour problems in pet dogs. Carlstead and colleagues (1999a,b) demonstrated both types of construct validity for keeper ratings of rhinoceros behaviour, while Wielebnowski (1999) did likewise for cheetahs. In both cases, discriminant validity was established by showing that the ratings differed between animals that differed in age, sex or breeding success, as could be predicted *a priori*. These and several other studies, in domesticated species as well as zoo animals, have tested convergent validity by comparing qualitative ratings to the results of standard behavioural tests such as novel object tests (e.g. horses: Le Scolan, 1997) or direct quantifications of specific behaviours where appropriate (e.g. cats: Feaver et al., 1986), and have found significant correlations. Piecing together the evidence from a range of studies, it appears that qualitative ratings, and more specifically, those conducted by people familiar with the individual animals, can be both reliable and valid.

Applying questionnaires

Cooperation and communication among researchers can play a key role in making qualitative ratings more useful. As mentioned above, having more than one person involved in the writing of a questionnaire is beneficial. Once it has been written and validated, it should be made available to others. Rather than taking the time to develop and analyse new questionnaires each time, researchers can then build on the work of those who have gone before. This will allow the

methods to be analysed more thoroughly and continually improved. Since questionnaires are likely to be developed by independent research groups rather than large governing bodies, in most cases, the process of refining them will depend on the initiative of individual scientists. Standardization and widespread use of a questionnaire also allows for clear comparisons of results across studies.

One barrier to comparison between studies is the normative nature of ratings of behaviour and temperament. The ratings will depend on the observer's range of experience. The provision of clear descriptions of what was considered a normal or extreme expression of a behaviour or phenomenon, as by Stevenson-Hinde and colleagues (1980), can limit this effect and enable other researchers to draw more informed conclusions from the work. Training of the observers can also be used to manipulate what is perceived as the norm, by ensuring that they have been exposed to as full a range of the species' behaviour as possible. It should be noted that training does not necessarily improve the accuracy of discriminations by observers (Renner & Renner, 1993). Wemelsfelder et al. (2000) did find good agreement on ratings of behavioural traits even by untrained, inexperienced observers; however, these ratings would not necessarily be comparable with ratings of pigs from different populations. For multi-institutional analyses with different observers at each institution, it would be preferable for them to have been exposed to a more representative cross-section of the species. To this end, it would be advisable to provide a training session at the outset of the study, including the use of videos with examples of what might be considered extremes or typical instances of a behaviour pattern for the species of interest. Of course, training is time-consuming. In some cases, such as for ratings of social status, relative ranking within the social group may prove more important than an absolute rating compared to average members of the species. For some studies, then, training may not be worthwhile, and providing descriptions or definitions of the terms may be sufficient. In other cases, however, training remains necessary to improve external validity.

Meta-analyses or very large-scale multi-institutional studies in which the cultural setting varies between institutions must consider systematic biases in the ratings from one group. Individuals can have biases toward using a certain portion of a rating scale; for example, one might always tend to give extreme ratings while another uses only the mid-range of the scale. These biases are influenced not only by individual traits but also by the observer's cultural background (Matsumoto & Juang, 2004). Individual differences will not create a systematic bias within a study, but a tendency for all observers in one study to use a certain portion of the scale

would decrease the validity of comparisons with other studies. Another source of systematic bias might be a desire to enhance self-image by giving ratings that would reflect positively on their care of the animals or on the institution where they work. In some cultures, people seem to be less biased towards positive self-evaluations (Kitayama et al., 1997), but might be equally or more favourably biased in their evaluations which reflect on their in-group, or in this case, employer (e.g. Sedikides et al., 2003). The former phenomenon could influence ratings by pet owners, while concern for group image might affect ratings that reflect on an entire institution such as a zoo; as a result, this psychological motivation to maintain a positive self-image would have different effects on the findings of multi-national studies depending on the subject of the study.

Only in analyses utilizing data from around the world are cross-cultural differences in self-promotion or scale usage a major concern. In other studies, this bias should at least be balanced among all treatments or institutions, since there is no *a priori* reason to expect systematic differences in people's motivation to promote a positive image of their treatment of animals within a given culture. For analyses of data from multiple cultures, there are several ways that this problem could be handled. As mentioned above, in some cases, it might be appropriate to look at relative rankings of animals within a social group. The data from each study could then be converted to ordinal rankings on a particular item or trait, eliminating the problems of between-study differences in scale usage. When relative measures are insufficient, but the mean ratings differ between primary studies for a known reason which is not relevant to the aims of the analysis, ratings can be standardized by dividing by the mean for that study. In any other situation, differences must be tested for and controlled or at least reported in meta-analyses. If there is concern of inter-institutional differences when conducting primary research, videos could be exchanged anonymously between institutions to be rated by one impartial observer. The problem with this method is that discrepancies between the ratings might indicate a bias in the ratings provided by the animal's caretakers, but might also result from the impartial observer's limited knowledge of the individual, since integration over long spans of time is a major advantage of using caretaker ratings. Each study will have to consider the optimal way of handling cross-cultural differences.

Limitations of the qualitative assessment method

One limitation of qualitative or subjective ratings, apart from concerns with their susceptibility to biases, is that they may not be equally effective for all animals and all types of behaviour. The validation studies described above found that reliability varied depending on the parameter being measured. Carlstead and colleagues (2000) found that inter-observer agreement also varied between zoo species. They posited that this might be due to variation in the amount of time the keepers spent with their animals; other potential reasons include the more limited behavioural repertoires of some species (Gosling, 2001) or the increased difficulty in interpreting the mood or motivations of animals more divergent from humans. If a species relies heavily on modes of communication that are outside humans' sensory capabilities, such as ultrasound, it is possible to miss important behavioural responses entirely. For this reason, it can never be assumed that observer ratings alone give a complete picture of the situation, and like most methods, they are best used in conjunction with other measures. In addition, reliability and validity of personality ratings are known to be lower in immature animals (Jones and Gosling, 2005). It may be that this method cannot be applied to every group of animals or every type of behaviour. Nonetheless, it seems to be useful for most species that have been tested so far, and no method can be applied in an identical manner to all living species.

Another difficulty lies in determining the appropriate standard for comparison of behaviour. The norm for behaviour in captive populations may, in fact, reflect a state of poor welfare; however, wild populations cannot always be used as the example of desirable behaviour, because their different environment may necessitate different responses. The most appropriate standard will likely depend on the aims of the research, but at present, is likely to be typical captive members of the species, or captive members with minimal obvious signs of dysfunction such as near constant stereotypy or inability to reproduce. As research sheds more light on species-specific indicators of positive welfare, they may be incorporated into this standard.

Conclusions

Despite the challenges and limitations, there are several advantages to the use of qualitative ratings by caretakers or handlers. From a practical standpoint, they can save a great deal of time and money by eliminating the need for researchers to spend hours doing formal observations; neither do they require special equipment as some standard behavioural tests would. For zoo biologists,

this is particularly important because to obtain adequate sample sizes for statistical power, and to disentangle the many variables that might affect the animals, it is necessary to observe animals at multiple, geographically separated institutions. Qualitative assessments also integrate information over a long period, allowing the assessment of long-term welfare, which can be difficult using physiological measures such as corticosteroid levels that vary constantly, unless the measures are repeated frequently. Standardization of direct behavioural observations across institutions is extremely difficult because conditions differ and all animals must be in the same state at the time of testing, including seasonal effects and reproductive status (Carlstead et al., 2000). In addition to extending scientific knowledge and thus being ultimately beneficial to animal welfare, this method has a more proximate animal welfare benefit. Standardized behavioural tests commonly involve exposing the animal to an anxiety-inducing situation; being observed by unfamiliar people is likewise believed to be frightening or stressful (zoos: Chamove et al., 1988; labs: Mason et al., 2004). When ratings are obtained from caretakers, this stress is avoided.

This method could be applied to research on a variety of topics. As mentioned earlier, it is widespread in studies of personality across species. Understanding of personality, by which I mean consistent individual styles of behaviour, is important to the study of animal welfare because the significance of behaviours used as welfare indicators can vary between individuals. Carlstead, an advocate of the method in the field of zoo biology, has recommended that keepers should be asked to monitor behaviour and record social and reproductive status of each animal as well as environmental conditions daily (personal communication, 2006). This would facilitate the detection of subtle changes in an individual over time that might indicate when an individual is ill or otherwise in need of special attention. Analysis of data provided by keepers at multiple institutions would also allow identification of both potential stressors and features of the environment that promote desirable behaviours. This could aid in solving breeding problems as well as improving welfare. Similar methods could be utilized for epidemiological studies of problem behaviours in companion or animals.

With rigorous methods including the use of multiple observers, qualitative or subjective assessments by people who are familiar with the individual animals can be a legitimate scientific tool. In fact, they often undergo more thorough validation procedures than do measures that appear more objective on the surface. Thus, qualitative measures that are formalized and become widespread may, in some cases, be more reliable and valid than the standard methods of behaviour

assessment. They possess the advantages of providing important information about animal behaviour and welfare without causing any extra distress to the study subjects and with minimal cost to the researchers. Because they allow integration of complex behaviour patterns over time, they provide some insight into the subjective states in non-human animals. For these reasons, caretaker or handler ratings could make an important contribution to animal welfare science.

Acknowledgements

Thanks to Kathy Carlstead for inspiration and Georgia Mason for constructive feedback on an earlier draft

References

Block, J. 1965. *The Challenge of Response Sets: Unconfounding Meaning Acquiescence, and Social Desirability in the MMPI*. New York: Appleton Century Crofts.

Carlstead, K., Fraser, J., Bennett, C., & Kleiman, D. G. 1999b. Black rhinoceros (*Diceros bicornis*) in US zoos: II. Behavior, breeding success, and mortality in relation to housing facilities. *Zoo Biology*, 18, 35-52.

Carlstead, K., Mellen, J., & Kleiman, D. G. 1999a. Black rhinoceros (*Diceros bicornis*) in US zoos: I. Individual behavior profiles and their relationship to breeding success. *Zoo Biology*, 18, 17-34.

Carlstead, K., Shepherdson, D., Sheppard, C., Mellen, J. and Bennet, C. 2000. *Constructing Behavioural Profiles for Zoo Animals: Incorporating Behavioural Information into Captive Population Management*. American Zoo and Aquarium Association's Behaviour and Husbandry Advisory Group and Oregon Zoo.
http://lpzoo.org/ethograms/MBA_Techniques_Manual.doc

Chamove, A. S., Hosey, G. R., & Schaetzel, P. 1988. Visitors excite primates in zoos. *Zoo Biology*, 7, 359-369.

Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Feaver, J., Mendl, M., & Bateson, P. 1986. A method for rating the individual distinctiveness of domestic cats. *Animal Behaviour*, 34, 1016-1025.

Fraser, D. 1999. Animal ethics and animal welfare science: Bridging the two cultures. *Applied Animal Behaviour Science*, 65, 171-189.

Gold, K. C. & Maple, T. L. 1994. Personality-assessment in the gorilla and its utility as a management tool. *Zoo Biology*, 13, 509-522.

- Gosling, S. D.** 2001. From mice to men: What can we learn about personality from animal research? *Psychological Bulletin*, 127, 45-86.
- Hawkins, P.** 2002. Recognizing and assessing pain, suffering and distress in laboratory animals: A survey of current practice in the UK with recommendations. *Laboratory Animals*, 36, 378-395.
- Heidegger, M.** 1984. *Nietzsche: Volume II: The eternal recurrence of the same*. (Translation by D.F. Krell). New York: Harper. (Original work published 1937).
- Hewetson, M., Christley, R. M., Hunt, I. D., & Voute, L. C.** 2006. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. *Veterinary Record*, 158, 852-857.
- Hsu, Y. Y. & Serpell, J. A.** 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *Journal of the American Veterinary Medical Association*, 223, 1293-1300.
- Jones, A. C. & Gosling, S. D.** 2005. Temperament and personality in dogs (*Canis familiaris*): A review and evaluation of past research. *Applied Animal Behaviour Science*, 95, 1-53.
- Keeley, B. L.** 2004. Anthropomorphism, primatomorphism, mammalomorphism: Understanding cross-species comparisons. *Biology & Philosophy*, 19, 521-540.
- Kitayama, S., Markus, H. R., Matsumoto, H., & Norasakkunkit, V.** 1997. Individual and collective processes in the construction of the self: Self-enhancement in the united states and self-criticism in japan. *Journal of Personality and Social Psychology*, 72, 1245-1267.
- Lehman, H.** 1992. In *The Inevitable Bond: Examining Scientist-Animal Interactions* (Ed. by H. Davis), pp. 383-396. New York: Cambridge University Press.
- LeScolan, N., Hausberger, M., & Wolff, A.** 1997. Stability over situations in temperamental traits of horses as revealed by experimental and scoring approaches. *Behavioural Processes*, 41, 257-266.
- Martin, P. & Bateson, P.** 1993. *Measuring Behaviour: An Introductory Guide*, 2nd ed. Cambridge University Press. Cambridge, UK.
- Matsumoto, D. & Juang, L.** 2004. *Culture and Psychology*. Toronto, ON: Thomson Learning – Wadsworth.
- Mason, G., Wilson, D., Hampton, C., & Wurbel, H.** 2004. Non-invasively assessing disturbance and stress in laboratory rats by scoring chromodacryorrhoea. *Atla-Alternatives to Laboratory Animals*, 32, 153-159.

- Millspaugh, J. J., Woods, R. J., Hunt, K. E., Raedeke, K. J., Brundige, G. C., Washburn, B. E., & Wasser, S. K.** 2001. Fecal glucocorticoid assays and the physiological stress response in elk. *Wildlife Society Bulletin*, 29, 899-907.
- Panksepp, J.** 2003. Can anthropomorphic analyses of separation cries in other animals inform us about the emotional nature of social loss in humans? *Psychological Review*, 110, 376-388.
- Renner, M. J. & Renner, C. H.** 1993. Expert and novice intuitive judgments about animal behavior. *Bulletin of the Psychonomic Society*, 31, 551-552.
- Rousing, T. & Wemelsfelder, F.** 2006. Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. *Applied Animal Behaviour Science*, 101, 40-53.
- Rushen, J.** 1991. Problems associated with the interpretation of physiological data in the assessment of animal welfare. *Applied Animal Behaviour Science*, 28, 381-386.
- Saxena, S. & Orley, J.** 1997. Quality of life assessment: The world health organization perspective. *European Psychiatry*, 12, S263-S266.
- Schilhab, T. S. S.** 2002. Anthropomorphism and mental state attribution. *Animal Behaviour*, 63, 1021-1026.
- Sedikides, C., Gaertner, L., & Toguchi, Y.** 2003. Pancultural self-enhancement. *Journal of Personality and Social Psychology*, 84, 60-79.
- Serpell, J. A. & Hsu, Y. Y.** 2001. Development and validation of a novel method for evaluating behavior and temperament in guide dogs. *Applied Animal Behaviour Science*, 72, 347-364.
- Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M.** 1980. Subjective assessment of rhesus monkeys over four successive years. *Primates*, 21, 66-82.
- Tyler, T.** 2003. If horses had hands. *Society & Animals*, 11, 267-281.
- Vanderwolf, C. H.** 1998. Brain, behavior, and mind: What do we know and what can we know? *Neuroscience and Biobehavioral Reviews*, 22, 125-142.
- Wemelsfelder, F.** 1997. The scientific validity of subjective concepts in models of animal welfare. *Applied Animal Behaviour Science*, 53, 75-88.
- Wemelsfelder, F., Hunter, E. A., Mendl, M. T., & Lawrence, A. B.** 2000. The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science*, 67, 193-215.
- Wemelsfelder, F., Hunter, T. E. A., Mendl, M. T., & Lawrence, A. B.** 2001. Assessing the 'whole animal': A free choice profiling approach. *Animal Behaviour*, 62, 209-220.

Wielebnowski, N. C. 1999. Behavioral differences as predictors of breeding status in captive cheetahs. *Zoo Biology*, 18, 335-349.

Wright, J. G. & Feinstein, A. R. 1992. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating-scales. *Journal of Clinical Epidemiology*, 45, 1201-1218.