JZAR JOURNAL OF ZOO AND AQUARIUM RESEARCH
www.jzar.org



Research article

# Development of an animal welfare tool for zoo-housed primates, considering intra- and inter-rater reliability

**Maeva Primault[1,2] and Arnaud Dazord[1]**

[1]*Parc zoologique de la Bourbansais, 35720 Pleugueneuc, France*
[2]*Institut national supérieur des sciences agronomiques, de l'alimentation et de l'environnement, 21000 Dijon, France*

*Correspondence: Maeva Primault, email; maeva.primault@gmail.com*

**Keywords:** assessment, inter-rater reliability, intra-rater reliability, primates, tool, welfare, zoo

**Article history:**
Received: 13 Sept 2021
Accepted: 21 Mar 2023
Published online: 30 Apr 2024

## Abstract

Over the past few decades, the concept of animal welfare has become increasingly prominent in Western societies. Faced with this collective awareness, various organisations are now making it a point of honour to consider the welfare of their animals. This highlights the need for an objective welfare assessment that endeavours to understand animals' real feelings. Although consensus tools have been developed for livestock, there is still a long road ahead for zoo animals. The aim of this study, conducted at the Bourbansais Zoo, was to create a welfare assessment protocol specifically for zoo-housed primates. It was developed using: (i) the Five Domains model; (ii) indicators proposed in various animal welfare assessment protocols for zoo species and (iii) information on primate biology and welfare from the literature. Intra- and inter-observer reliability was evaluated using Gwet's AC1 coefficient. A total of 21 of the 29 criteria in the evaluation grid are characterised by 'good' to 'very good' inter-rater reliability, according to the Altman classification. The results also reveal promising intra-observer reliability with only three criteria having an AC1 score below 0.6. Given these satisfactory results, the tool could prove to be useful for monitoring primate wellbeing over time, implementing corrective actions and evaluating the effects of these modifications.

## Introduction

Over the past thirty years, there has been growing collective awareness of the concept of animal welfare, which is becoming an increasingly important issue in Western societies. Citizens are now actively participating in debates on the subject, motivated by better visibility of the living conditions of farm animals and by changes in people's relationship with animals (Delanoue et al. 2018).

Faced with these new societal demands, it is essential that public and private institutions take into account the welfare of their animals. The question therefore arises of an objective assessment of animal welfare that takes into account animals' behavioural signals, without falling into anthropomorphism (Green and Mellor 2011; Nuffield Council on Bioethics 2005; Whitham and Wielebnowski 2013). Protocols such as Welfare Quality®, a programme funded by the European Commission and involving 44 institutions from 17 countries, have been developed to assess the welfare of farm animals (Welfare Quality® 2007). Initially designed for well-known domestic species such as cows, pigs and poultry (Welfare Quality® 2009a, b, c), these protocols are not always suitable for assessing captive wild animals.

Zoos, like livestock farms, are also affected by these changes in attitudes to animal welfare (Roe et al. 2014; Sherwen et al. 2018; Whitham and Wielebnowski 2013). Even though the development of a consensus and functional tools remains a real challenge (Hill and Broom 2009; Melfi 2005) due to the large number of extremely different animal species held, more and more zoo institutions are developing their own welfare

assessment methods to: (i) monitor the condition of their animals; (ii) identify priority areas for improvement and (iii) meet societal expectations.

The Five Domains model, originally developed by Mellor and Reid (1994), is now considered the most appropriate framework for assessing animal welfare. It is recommended by the World Association of Zoos and Aquariums (WAZA), the Royal Society for the Prevention of Cruelty to Animals (RSPCA) and the British and Irish Association of Zoos and Aquariums (BIAZA) for zoos (BIAZA 2020; RSPCA 2013; WAZA 2015). It is a tool to reliably and systematically assess the level of an animal's welfare at a given time, and as far as possible to determine an animal's real feelings (Mellor 2017). The first four domains of the model (nutrition, physical environment, health and behaviours) draw attention to the negative impact that external factors can have on organism functions. The last domain (mental state) was designed to capture the overall mental experience of the animals due to all the impacts considered in the first four domains. Factors in the physical and functional domains (domains 1 to 4) have emotional consequences, or "affects", that are subjective and internalised by the animal and therefore elusive to humans. All these affects are then attributed to the fifth domain (Mellor 2017; Mellor and Beausoleil 2015). For example, tissue injury (domain 3) stimulates nociceptors which propagate neural impulses to the brain where they are translated into pain experience (domain 5) (Dubin and Patapoutian 2010). Overall, the balance between positive and negative affects plays a key role in the animal's welfare. The Five Domains model thus allows for an assessment of animal welfare (Hemsworth et al. 2015; Mellor 2016, 2017; Mellor and Beausoleil 2015). In addition, similarly to the Five Freedoms model (Department of Environment, Food and Rural Affairs and Farm Animal Welfare Council 1993; Webster 1995, 2005) that previously proved successful by incorporating for the first time the notion of subjective experience, the latest version of Mellor's model includes a special focus on the animal's positive experiences (Mellor et al. 2020), the importance of which has been widely demonstrated in welfare assessments since the 2000s (Green and Mellor 2011; Lawrence et al. 2019; Mellor 2016; Yeates and Main 2008).

This study, conducted at the Bourbansais Zoo in Brittany, France, aimed to develop a welfare assessment tool specifically dedicated to zoo-housed primates, using the Five Domains Model. Although primates are largely represented in zoos around the world, there is no non-invasive welfare assessment protocol specifically designed for them. The main purpose of this work was to field-test the tool, focusing on two reliability parameters: inter- and intra-rater reliability.

## Materials and methods

### *Development of welfare assessment tool*

The assessment tool is based on the operational details of the Five Domains model (Mellor 2017; Mellor et al. 2020), a brief review of the indicators suggested in a variety of published species-specific zoo animal welfare assessment tools (Justice et al. 2017; Rose and O'Brien 2020; Wildlife Reserves Singapore n.d.; Yon et al. 2019) and information on primate welfare drawn from published literature (Bassett and Buchanan-Smith 2007; Buchanan-Smith et al. 2002; Canadian Council on Animal Care 2019; Clingerman and Summers 2012; Colleen et al. 2007; Department for Environment, Food and Rural Affairs 2010; Global Federation of Animal Sanctuaries 2013a, b; Jennings et al. 2009; Line et al. 1991; Lutz and Novak 2005; Novak et al. 2012; Pomerantz et al. 2013; Rennie and Buchanan-Smith 2006; Schmidt 2011; Weiss and Hampshire 2015; Wolfensohn et al. 2018).

The tool consists of four main parts, matching the first four domains of the Five Domains model (Mellor and Beausoleil 2015): nutrition, physical environment, health, and behaviours and behavioural interactions. The last domain was divided into four sub-categories adapted from those proposed in the latest version of the Five Domains model (Mellor et al. 2020): behavioural interactions with the environment, with other non-human animals, with humans and individual behaviours. In each domain, criteria were defined, providing a total of 29 criteria. Criteria that could be classified in several domains were assigned to only one in order not to count the same information twice. As recommended in the literature, the grid includes management-related criteria related primarily to health and feeding, resource-based criteria, and physical and behavioural animal-based criteria (Manteca Vilanova 2020; Sherwen et al. 2018; Whitham and Wielebnowski 2013; Wolfensohn et al. 2015). Special attention was paid to incorporating positive behavioural indicators, and not being restricted to eliminating negative behaviours, in line with the Five Domains model (Manteca Vilanova 2020; Mellor and Beausoleil 2015). The tool is intended to be tested at the individual level, which is the ideal level for addressing animal welfare (Barber 2009).

The grid is characterised by a four-level scoring system, the meaning of which is detailed for each criterion, to limit subjective interpretation as much as possible. The option 'not applicable' was added for cases where the criterion does not apply to the animal and for specific situations in which a score cannot be chosen (see Supplementary Information). For greater convenience during evaluation sessions, the grid was converted into a Google Forms questionnaire to be completed directly online by the raters.

In order to provide guidance to users, three appendices were developed. The first two are used to assign the Body Condition Score (BCS) and Coat Condition Score for criteria 5 and 18. The third provides a definition of each behaviour or activity present in the grid, which is particularly useful when dealing with criteria 24, 28 and 29 (see Supplementary Information).

### *Subjects*

The welfare assessment tool was tested at the Bourbansais Zoo, where 28 individuals from 14 primate species were assessed. These animals were selected randomly, with two individuals per species (see Table 1).

All these primates live in groups whose composition depends on the species, in naturalistic outdoor enclosures. Ring-tailed lemurs *Lemur catta* (three males), *Eulemur coronatus* crowned lemurs (three males and two females), red-ruffed lemurs *Varecia rubra* (two males) and red-bellied lemurs *Eulemur rubriventer* (two males) live together in a 'walk-through' enclosure. The pair of lar gibbons *Hylobates lar* lives in a closed-roof enclosure, as do the pair of pygmy marmosets *Cebuella pygmaea* and the group of silvery marmosets *Mico argentatus* (two males) and emperor tamarins *Saguinus imperator* (two females). The tufted capuchins *Sapajus apella* (two males and one female), black-capped squirrel monkeys *Saimiri boliviensis* (seven males), black-headed spider monkeys *Ateles fusciceps* (three males and five females), mantled guerezas *Colobus guereza* (three males) and diana monkeys *Cercopithecus diana* (one male and two females) each live on their respective islands. Finally, the two male geladas *Theropithecus gelada* live in a 'classic' outdoor enclosure. They all benefit from free access to indoor housing during daylight hours.

### *Data collection*

The assessment sessions were conducted on one to two animals per day at 1400 Monday to Friday, for a total duration of 6 weeks from 17 May to 25 June 2021.

Inter-rater reliability refers to the agreement between several experimenters independently and simultaneously rating the same individual (Meagher 2009; Yon et al. 2019). To assess inter-rater

**Table 1.** List of assessed individuals

| Family | Species | ID | Sex | Birth year |
|---|---|---|---|---|
| Lemuridae | *Lemur catta* | LC1 | M | 2009 |
| | | LC2 | M | 2002 |
| | *Eulemur rubriventer* | ER1 | M | 2012 |
| | | ER2 | M | 2011 |
| | *Eulemur coronatus* | EC1 | M | 2019 |
| | | EC2 | F | 2013 |
| | *Varecia rubra* | VR1 | M | 2000 * |
| | | VR2 | M | 2000 |
| Callitrichidae | *Mico argentatus* | MA1 | M | 2018 * |
| | | MA2 | F | 2018 |
| | *Saguinus imperator* | SI1 | F | 2011 |
| | | SI2 | F | 2010 |
| | *Cebuella pygmaea* | CP1 | M | 2017 |
| | | CP2 | F | 2015 |
| Cebidae | *Sapajus apella* | SA1 | M | 2003 |
| | | SA2 | F | 2004 |
| | *Saimiri boliviensis* | SB1 | M | 2000 |
| | | SB2 | M | 2014 * |
| Atelidae | *Ateles fusciceps* | AF1 | F | 2000 * |
| | | AF2 | M | 2016 |
| Cercopithecidae | *Colobus guereza* | CG1 | M | 1999 |
| | | CG2 | M | 2002 |
| | *Cercopithecus diana* | CD1 | F | 2017 * |
| | | CD2 | M | 2003 |
| | *Theropithecus gelada* | TG1 | M | 2011 |
| | | TG2 | M | 2012 |
| Hylobatidae | *Hylobates lar* | HL1 | F | 1988 |
| | | HL2 | M | 1990 * |

* Animals assessed three days in a row

reliability, three observers—the first author, the zoo manager and a zookeeper—carried out the assessment simultaneously every day for approximately 20 minutes. Applied to the welfare assessment, intra-rater reliability refers to the agreement between ratings made by the same individual over a short period of time, often between two and seven days (Harvey 2021; Meagher 2009; Yon et al. 2019). To evaluate intra-rater reliability, one individual was selected at random from each primate family and the assessment was conducted on Tuesday, Thursday and Friday of each week by the zoo manager and the first author, following the same protocol as presented above (see Table 1).

The evaluation sessions took place outside the enclosure in order to reduce the impact of the raters on the focal animal. However, assessors were allowed to enter the enclosure to assess physical animal-based criteria (e.g. BCS, coat condition, signs of diseases) and to inspect the indoor housing and other areas of the enclosure to assess criteria relevant to them. Throughout the assessment, the raters had access to printouts of the evaluation grid and its appendices. Raters were asked not to submit the Google Form before the end of the 20-minute observation session, so that they could revise their answers if necessary.

***Statistical analysis***

The scores obtained per criterion and per rater at each session were reported on a Microsoft Excel spreadsheet so that they could be processed using the statistical computing language R, version 4.0.3.

Intra- and inter-rater reliabilities, in the case of categorical data, are mostly evaluated using Cohen's kappa. This measures agreement between observers, based on the ratio of observed

**Table 2.** Strength of reliability according to Altman's classification (Altman 1990)

| Gwet's AC1 value | Strength of agreement (Altman 1990) |
|---|---|
| <0.2 | Poor |
| 0.21–0.4 | Fair |
| 0.41–0.6 | Moderate |
| 0.61–0.8 | Good |
| 0.81–1 | Very good |

agreement between raters to the probability of a random agreement (Giammarino et al. 2021). However, studies have shown that this coefficient has limitations, especially when rating asymmetries, i.e. when one or several rating level(s) are largely over-represented. In this case, even if the raters sometimes completely agree, kappa might be interpreted as neutral or even negative agreement (Wongpakaran et al. 2013; Zec et al. 2017). To overcome this problem, an alternative coefficient was developed by Gwet in 2014, called the first-order agreement coefficient or AC1, which adjusts for chance agreement (Gwet 2019; Wongpakaran et al. 2013). This coefficient can be used for two or more assessors, provided that a categorical rating system is used (Gwet 2019; Wongpakaran et al. 2013).

Gwet's AC1 was therefore used to assess the intra- and inter-observer reliability of each criterion of the evaluation grid using the statistical processing software R. The values of the Gwet AC1 coefficients were interpreted according to the reliability levels presented in Table 2.

Average reliabilities per domain (nutrition, physical environment, health, behaviour and behavioural interactions) and per type of criteria (animal-based, resource-based or management-based) were calculated by averaging the AC1 coefficients of the corresponding criteria, together with the corresponding standard deviations.

## Results

### *Inter-rater reliability*

The Gwet AC1 coefficients obtained for inter-observer reliability of the 29 criteria of the grid are presented in Table 3. They take into account all the observers and individuals assessed.

Criteria 1, 2 and 19, respectively relating to food preparation, food rations and the animal's ease of movement, have the highest AC1 values (total agreement; AC1=1). According to Altman's classification, 13 other criteria are characterised by very good reliability (AC1 between 0.819 and 0.976). These are mainly animal-based criteria (16: diseases, 17: injuries, 18: coat condition, 24: social behaviours and 29: stereotypical individual behaviours) and resource-based criteria (6: cleanliness and security, 8: shelter, 9: temperature, 10: humidity, 11: light, 22: substrate and 27: isolation possibility).

In contrast, the criteria with the lowest AC1 values are criteria

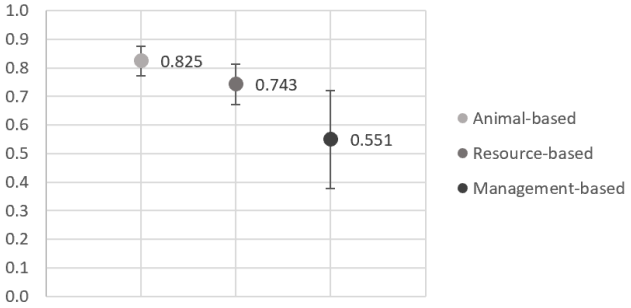


**Figure 1.** Gwet's AC1 mean and standard error per type of criterion

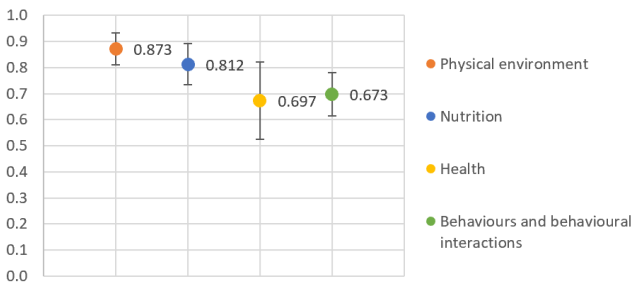


**Figure 2.** Gwet's AC1 mean and standard error per domain

**Table 3.** Gwet's AC1 values and strength of agreement per criterion

| Domain | Criterion | Type of criterion | Gwet's AC1 | Standard error | Strength of agreement (Altman 1990) |
|---|---|---|---|---|---|
| Nutrition | 1 | Management | 1 | 0 | Very good |
| | 2 | Management | 1 | 0 | Very good |
| | 3 | Management | 0.611 | 0.085 | Good |
| | 4 | Resource | 0.701 | 0.102 | Good |
| | 5 | Animal | 0.774 | 0.074 | Good |
| Physical Environment | 6 | Resource | 0.976 | 0.025 | Very good |
| | 7 | Resource | 0.565 | 0.126 | Moderate |
| | 8 | Resource | 0.976 | 0.025 | Very good |
| | 9 | Resource | 0.918 | 0.049 | Very good |
| | 10 | Resource | 0.976 | 0.025 | Very good |
| | 11 | Resource | 0.923 | 0.046 | Very good |
| | 12 | Resource | 0.562 | 0.121 | Moderate |
| Health | 13 | Management | 0.562 | 0.121 | Moderate |
| | 14 | Management | 0.09 | 0.086 | Poor |
| | 15 | Management | 0.04 | 0.137 | Poor |
| | 16 | Animal | 0.976 | 0.025 | Very good |
| | 17 | Animal | 0.975 | 0.026 | Very good |
| | 18 | Animal | 0.861 | 0.061 | Very good |
| | 19 | Animal | 1 | 0 | Very good |
| Behaviours and Behavioural Interactions | 20 | Resource | 0.77 | 0.083 | Good |
| | 21 | Resource | 0.176 | 0.084 | Poor |
| | 22 | Resource | 0.819 | 0.064 | Very good |
| | 23 | Resource | 0.377 | 0.073 | Fair |
| | 24 | Animal | 0.837 | 0.07 | Very good |
| | 25 | Animal | 0.473 | 0.102 | Moderate |
| | 26 | Animal | 0.665 | 0.097 | Good |
| | 27 | Resource | 0.923 | 0.046 | Very good |
| | 28 | Animal | 0.771 | 0.076 | Good |
| | | Animal | 0.918 | 0.049 | Very good |

14: veterinary examinations, 15: preventive medicine, 21: enrichments and 23: social structure (poor to low reliability; AC1 between 0.04 and 0.377). These four criteria correspond to indicators based on the environment or management.

In terms of trend, indicators with the best inter-observer mean reliability are those based on animal observations ($AC1_{mean}$=0.825) (Figure 1). The domain that appears to have the best inter-observer mean reliability is the physical environment ($AC1_{mean}$=0.842). The health domain ($AC1_{mean}$=0.643) has the lowest mean score and the most dispersed AC1 values (Figure 2).

***Intra-rater reliability***

The Gwet AC1 coefficients obtained for intra-observer reliability of the 29 criteria of the grid are presented in Table 4. They take into account all the observers and individuals evaluated.

In total, nine criteria have AC1 values equal to 1, reflecting the best intra-observer reliability. These are management-based criteria (1: food preparation, 2: ration and 3: food distribution), resource-based criteria (6: cleanliness and security, 8: shelter, 10: humidity and 27: isolation possibility) and animal-based criteria (16: diseases and 17: injuries).

According to Altman's classification, 12 other criteria have very good reliability (AC1 between 0.803 and 0.941), 9 of which are related to management or resources, and 3 are animal-based.

In contrast, the criteria with the lowest intra-observer reliability are the criteria referring to the relationship with visitors (26) and non-stereotypical individual behaviours (28) (poor to low reliability; AC1 between 0 and 0.378).

Figure 3 shows that in terms of trends, the type of indicators with the best average intra-observer reliability are those based on environmental observations ($AC1_{mean}$=0.916). In Figure 4 the domain that appears to have the best intra-observer mean

**Table 4.** Gwet's AC1 value and strength of agreement per criterion

| Domain | Criterion | Type of criterion | Gwet's AC1 | Standard error | Strength of agreement (Altman 1990) |
|---|---|---|---|---|---|
| Nutrition | 1 | Management | 1 | 0 | Very good |
| | 2 | Management | 1 | 0 | Very good |
| | 3 | Management | 1 | 0 | Very good |
| | 4 | Resource | 0.869 | 0.103 | Very good |
| | 5 | Animal | 0.778 | 0.149 | Good |
| Physical Environment | 6 | Resource | 1 | 0 | Very good |
| | 7 | Resource | 0.903 | 0.102 | Very good |
| | 8 | Resource | 1 | 0 | Very good |
| | 9 | Resource | 0.941 | 0.06 | Very good |
| | 10 | Resource | 1 | 0 | Very good |
| | 11 | Resource | 0.931 | 0.072 | Very good |
| | 12 | Resource | 0.916 | 0.09 | Very good |
| Health | 13 | Management | 0.6 | 0.195 | Moderate |
| | 14 | Management | 0.676 | 0.132 | Good |
| | 15 | Management | 0.803 | 0.12 | Very good |
| | 16 | Animal | 1 | 0 | Very good |
| | 17 | Animal | 1 | 0 | Very good |
| | 18 | Animal | 0.781 | 0.145 | Good |
| | 19 | Animal | 0.929 | 0.076 | Very good |
| Behaviours and Behavioural Interactions | 20 | Resource | 0.876 | 0.093 | Very good |
| | 21 | Resource | 0.636 | 0.155 | Good |
| | 22 | Resource | 0.941 | 0.062 | Very good |
| | 23 | Resource | 0.889 | 0.111 | Very good |
| | 24 | Animal | 0.929 | 0.076 | Very good |
| | 25 | Animal | 0.778 | 0.15 | Good |
| | 26 | Animal | 0 | 0.288 | Poor |
| | 27 | Resource | 1 | 0 | Very good |
| | 28 | Animal | 0.378 | 0.225 | Fair |
| | 29 | Animal | 0.941 | 0.06 | Very good |

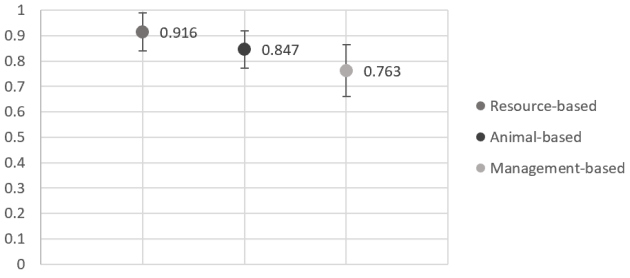


**Figure 3.** Gwet's AC1 mean and standard error per type of criterion

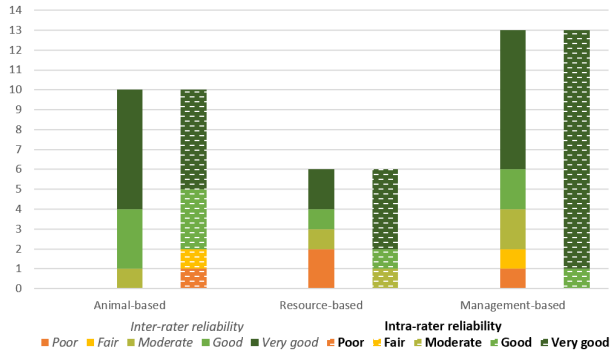


**Figure 4.** Gwet's AC1 mean and standard error per domain

reliability is the physical environment ($AC1_{mean}$=0.955) with a very low dispersion around the mean value. The behaviours and behavioural interactions domain ($AC1_{mean}$=0.749) has the lowest mean score and the most dispersed AC1 values.

In summary, animal-based criteria as well as nutrition and physical criteria seem to be characterised by the best inter-rater reliability, while management-based criteria are less satisfactory. Resource-based and management-based criteria seem to have the best intra-rater reliability, as do nutrition, physical environment and health criteria. Considering the two types of reliability, the criteria belonging to the behaviours and behavioural interactions domain appear to be the most challenging to score consistently (Figures 5 and 6).

## Discussion

### *Inter-rater reliability*

Among the criteria with the highest rater agreement, four health-related indicators stand out (ease of movement, injuries, diseases and coat condition). The high inter-observer reliability of these indicators corroborates the results of similar studies on horses and sheep, in which indicators chosen to assess good health are generally very reliable (Brule-Aupiais et al. 2015; Dany et al. 2017). Similarly, the very good scores obtained for behavioural criteria such as social or individual stereotypical behaviours are close to those obtained for horses (Dany et al. 2017) and for pigs (Velarde et al. 2007).

Yet several criteria seem to be characterised by low inter-observer reliability, including veterinary examinations and preventive medicine. This can be explained by the fact that the raters did not systematically have all the necessary information to objectively select a score level. Only the zoo manager knew all the veterinary practices, the frequency of visits by the referring veterinarian and the list of preventive treatments administered to each animal. For the object enrichments criterion, the low reliability is probably related to difficulties with the strict definition of the term 'enrichment' (Lutz and Novak 2005; Mellen and MacPhee 2001). Raters reported a variety of definitions, and objects identified by some as enrichments were not identified by others, despite the presence of examples in the evaluation grid. The low reliability of the social structures criterion is more surprising, because caretakers are expected to be familiar with the social structures typically observed in the wild for each species, as an integral part of their training. However, as social structures remain a complex notion, it is possible that mistakes were made by some raters.

Finally, it is interesting to note that the animal-based criteria have AC1 coefficients greater than 0.6, with the exception of criterion 25 on the relationship to caretakers, indicating good to very good inter-observer reliability (Altman 1990; Table 2). The results are all the more encouraging as these criteria are known to be more prone to subjectivity (Richmond et al. 2017; Vieira et al. 2018).

### *Intra-rater reliability*

Most of the criteria evaluated seem to have very good intra-observer reliability. Those with the highest AC1 are criteria that appear to be fairly fixed over time and therefore varied little over the duration of the intra-observer experiment, the total duration of which was only four days. It follows that the resource-based criteria were, on average, the most reproducible.

The criteria with the lowest reliability were animal-based and concerned the relationship with visitors and non-stereotypical individual behaviours. These low values can be explained by the presence of external biases. First, depending on the days chosen for the experiments, the number of visitors in the zoo was not fixed. It should be noted that on busy days, particularly when active groups such as schools were present, the animals were more stimulated and therefore exhibited more demonstrative behaviour—either positive or negative—towards visitors (Hashmi

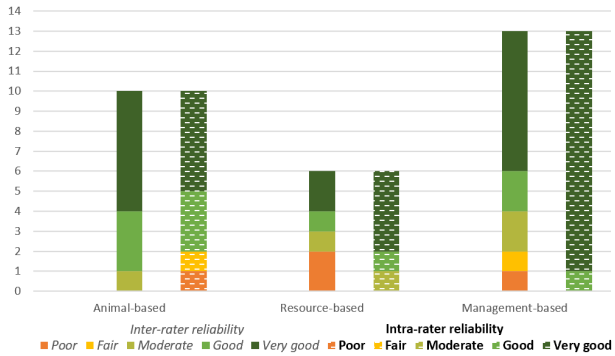


**Figure 5,** Strength of agreement (Altman 1990) by type of criterion comparing inter- and intra-rater reliability

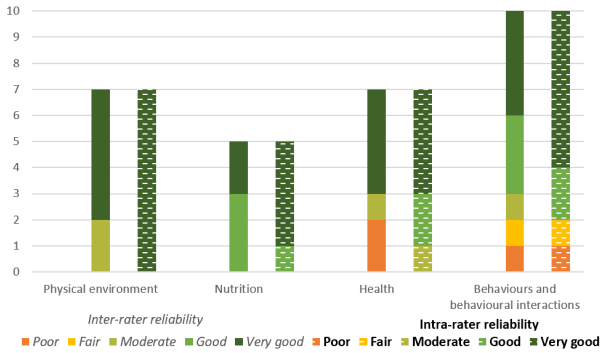


**Figure 6.** Strength of agreement (Altman 1990) in each domain comparing inter- and intra-rater reliability

and Sullivan 2020a; Mitchell et al. 1992; Wells 2005). On the other hand, when admissions were lower, animals were more likely to show indifference or even avoidance, as highlighted in other studies (Davey 2007; Hashmi and Sullivan 2020b; Sherwen and Hemsworth 2019). In addition, the presence and characteristics of visitors also influenced individual primate behaviour, for example by reducing foraging or playing patterns when visitor group size or induced noises increase as shown in previous studies (Fernandez et al. 2009; Wells 2005; Wood 1998).

Individual primate behaviours may have varied considerably due to changing weather conditions between observation sessions. For example, on rainy days the animals observed were generally sheltered in the trees or in their lodges and did not express a wide variety of positive individual behaviours. In such instances the raters did not assign the maximum score. In addition, periods of napping or resting, which are more frequent in rainy weather for lemurs in particular (Collins et al. 2017; Goodenough et al. 2019), could be interpreted as a kind of apathy and thus as negative individual behaviour, causing a lower score to be awarded.

***Other considerations***

The tool introduced in this paper is meant to be a comprehensive, practical, field-friendly tool. The fact that the grid can quickly be completed in about 20 minutes per enclosure contributes considerably to this. In addition, the criteria based on resources and management can be completed at another time, which further reduces the actual observation time. Given caretakers' often busy schedule, this is important. However, testing the protocol in real-life conditions showed that a 20-minute window at fixed times was not necessarily ideal, especially when the animal confined itself to a single behaviour during the entire observation period. To address this, shorter observation sessions at different times of the day could be set up, as proposed in other studies (Rose and O'Brien 2020; Yon et al. 2019).

Furthermore, one of the main purposes of this tool is for internal use by the caretakers themselves. This has a number of advantages, both from an organisational point of view and because the keepers know their animals very well; they are more sensitive to even slight behavioural changes and are therefore more likely to detect possible problems (Beaver and Bayne 2014; Whitham and Wielebnowski 2009). However, this may also bias the results with more positive ratings as caretakers may tend to believe that the level of wellbeing reflects the quality of care they provide to their animals (Sherwen et al. 2018). The importance of training caretakers in this type of assessment and of communicating its real objectives is thus highlighted.

The assessment carried out in this study focuses only on reproducibility (or, in statistical terms, 'precision'), leaving aside validity parameters such as concurrent criteria validity (or, in statistical terms, 'accuracy'). It is therefore impossible to know whether all assessors have agreed on a welfare level that reflects reality. Concurrent criteria validity is generally assessed by comparing results with independent 'gold standard' measures (Meagher 2009; Yon et al. 2019). As real levels of animal welfare are inherently unknown in such protocols, this validity parameter is often difficult to evaluate. In a future study it might be useful to compare the results obtained within caretakers' groups with those obtained by animal welfare experts and specialists, in order to further characterise the accuracy of the conclusions (Brouwers and Duchateau 2021; Sherwen et al. 2018; Yon et al. 2019). These tests were carried out on a modest sample of animals living in a limited range of environments. The tool has not been tested on the Lorisiformes and Tarsiiformes infra-orders, for example. Thus, to obtain more precise results and to analyse the reproducibility of the criteria in greater detail, it would be useful to run the tests in other zoos and on other species.

## Conclusion

For the first time, a protocol for assessing the welfare of primates in zoos was constructed based on the Five Domains model, which is the model currently recommended by various zoological institutions. This protocol was then tested on a sample of lemurs and monkeys from the Bourbansais Zoo. This study reveals promising results regarding the protocol's level of intra- and inter-observer reliability. In particular, some indicators showed excellent intra- and inter-observer reliability and thus seem very robust. This is the case not only for criteria based on the environment (e.g. cleanliness and security, shelter, abiotic parameters), but also for criteria based on the animal (e.g. ease of movement, injuries, diseases, behavioural indicators), which is very encouraging for the more regular introduction of this type of criteria in future welfare assessment protocols. Nevertheless, the results also show the importance of retaining environmental indicators, which are complementary to the animal-centred criteria.

While reliability parameters seem satisfactory, validity aspects need to be investigated in order to obtain an operational and scientifically sound protocol. In its fully functional form, the protocol could then become an interesting tool for zookeepers, allowing them to analyse primate wellbeing more precisely. In the future, it could additionally become a discussion tool to implement appropriate actions, based on the results obtained. The systematic application of the protocol over several years would allow assessment of whether adopted corrective measures had been effective by comparing the different scores, and thus would contribute to the principle of continuous improvement.

## Acknowledgments

The authors sincerely thank the staff of the Bourbansais Zoo for allowing the authors to carry out this study on their animals, for their participation in the experiments and for their warm welcome.

## References

Altman D.G. (1990) *Practical Statistics for Medical Research*. London, UK: CRC Press.

Barber J.C.E. (2009) Programmatic approaches to assessing and improving animal welfare in zoos and aquariums. *Zoo Biology* 28(6): 519–530. doi:10.1002/zoo.20260

Bassett L., Buchanan-Smith H.M. (2007) Effects of predictability on the welfare of captive animals. *Applied Animal Behaviour Science* 102(3–4): 223–245. doi:10.1016/j.applanim.2006.05.029

Beaver B.V., Bayne K. (2014) Animal welfare assessment considerations. In: Bayne K., Turner P.V. (eds.). *Laboratory Animal Welfare*. London, UK: Academic Press, 29–38.

BIAZA (2020) *BIAZA Animal Welfare Policy 2020*. London, UK: BIAZA. Available online at https://biaza.org.uk/policies-guidelines (accessed 13 April 2021).

Brouwers S., Duchateau M.J. (2021) Feasibility and validity of the Animal Welfare Assessment Grid to monitor the welfare of zoo-housed gorillas *Gorilla gorilla gorilla. Journal of Zoo and Aquarium Research* 9(4): 208–217. doi:10.19227/jzar.v9i4.607

Brule-Aupiais A., Mialon M.M., Gautier D., Pottier É., Ribaud D., Boissy A., Boivin X. (2015) Validation d'une méthode d'évaluation du bien-être des ovins en ferme et comparaison de deux types de conduites hivernales. *Presented at the 22 Rencontres autour des Recherches sur les Ruminants*, Institut de l'Elevage, Paris.

Buchanan-Smith H.M., Shand C., Morris K. (2002) Cage use and feeding height preferences of captive common marmosets (*Callithrix j. jacchus*) in two-tier cages. *Journal of Applied Animal Welfare Science* 5(2): 139–149. doi:10.1207/S15327604JAWS0502_045327604JAWS0502_04

Canadian Council on Animal Care (2019) *CCAC Guidelines: Nonhuman Primates.* Ottawa, Canada: Canadian Council on Animal Care.

Clingerman K.J., Summers L. (2012) Validation of a body condition scoring system in Rhesus macaques (*Macaca mulatta*): Inter- and intrarater variability. *Journal of the American Association for Laboratory Animal Science* 51(1): 31–36.

Colleen M., Buchanan-Smith H., Farmer K.H., Fitch-Snyder H., Lisa J.E., Prescott M., Sylvia T. (2007) *IPS International Guidelines for the Acquisition, Care and Breeding of Non-Human Primates*, Second ed, Primate Report — Special Issue. Bronx, New York: International Primatological Society.

Collins C., Corkery I., Haigh A., McKeown S., Quirke T., O'Riordan R. (2017) The effects of environmental and visitor variables on the behavior of free-ranging ring-tailed lemurs (*Lemur catta*) in captivity. *Zoo Biology* 36(4): 250–260. doi:10.1002/zoo.21370

Dany P., Vidament M., Yvon J.M., Reigner F., Barrière P., Riou M., Layne A.L., Lansade L., Minero M., dalla Costa E., Briant C. (2017) Protocole d'évaluation du bien-être chez le cheval "AWIN Horse" : essai en conditions expérimentales et premières évaluations sur le terrain. *43ème Journée de la Recherche Équine.* Paris, France: Institut Français du Cheval et de l'Equitation, 159–162.

Davey G. (2007) Visitors' effects on the welfare of animals in the zoo: A review. *Journal of Applied Animal Welfare Science* 10(2): 169–183. doi:10.1080/10888700701313595

Delanoue E., Dockes A.C., Chouteau A., Roguet C., Philibert A. (2018) Regards croisés entre éleveurs et citoyens français : vision des citoyens sur l'élevage et point de vue des éleveurs sur leur perception par la société. *INRAE Productions Animales* 31(1): 51–68. doi:10.20870/productions-animales.2018.31.1.2203

Department for Environment, Food and Rural Affairs (2010) *Code of Practice for the Welfare of Privately Kept Non-Human Primates*. London, UK: Department for Environment, Food and Rural Affairs.

Department of Environment, Food and Rural Affairs, Farm Animal Welfare Council (1993) *Second Report on Priorities for Research and Development in Farm Animal Welfare.* London, UK: Department of Environment, Food and Rural Affairs & Farm Animal Welfare Council.

Dubin A.E., Patapoutian A. (2010) Nociceptors: The sensors of the pain pathway. *Journal of Clinical Investigation* 120: 3760–3772.

Fernandez E.J., Tamborski M.A., Pickens S.R., Timberlake W. (2009) Animal–visitor interactions in the modern zoo: Conflicts and interventions. *Applied Animal Behaviour Science* 120(1–2): 1–8. doi:10.1016/j.applanim.2009.06.002

Giammarino M., Mattiello S., Battini M., Quatto P., Battaglini L.M., Vieira A.C.L., Stilwell G., Renna M. (2021) Evaluation of inter-observer reliability of animal welfare indicators: Which is the best index to use? *Animals* 11(5): 1445. doi:10.3390/ani11051445

Global Federation of Animal Sanctuaries (2013a) *Standards For Old World Primates*. Washington, DC: Global Federation of Animal Sanctuaries.

Global Federation of Animal Sanctuaries (2013b) *Standards For New World Primates*. Washington, DC: Global Federation of Animal Sanctuaries.

Goodenough A.E., McDonald K., Moody K., Wheeler C. (2019) Are "visitor effects" overestimated? Behaviour in captive lemurs is mainly driven by co-variation with time and weather. *Journal of Zoo and Aquarium Research* 7(2): 59–66. doi:10.19227/jzar.v7i2.343

Green T.C., Mellor D.J. (2011) Extending ideas about animal welfare assessment to include 'quality of life' and related concepts. *New Zealand Veterinary Journal* 59(6): 263–271.

Gwet K.L. (2019) *Computing Agreement Coefficients from Contingency Tables*. Available online at https://cran.r-project.org/web/packages/irrCAC/vignettes/overview.html (accessed 2 July 2021).

Harvey N.D. (2021) A simple guide to inter-rater, intra-rater and test-retest reliability for animal behaviour studies. *OSF Preprints*. doi:10.31219/osf.io/8stpy

Hashmi A., Sullivan M. (2020a) The visitor effect in zoo-housed apes: The variable effect on behaviour of visitor number and noise. *Journal of Zoo and Aquarium Research* 8: 268–282. doi:10.19227/jzar.v8i4.523

Hashmi A., Sullivan M. (2020b) The visitor effect in zoo-housed apes: The variable effect on behaviour of visitor number and noise. *Journal of Zoo and Aquarium Research* 8(4): 268–282. doi:10.19227/jzar.v8i4.523

Hemsworth P.H., Mellor D.J., Cronin G.M., Tilbrook A.J. (2015) Scientific assessment of animal welfare. *New Zealand Veterinary Journal* 63(1): 24–30. doi:10.1080/00480169.2014.966167

Hill S.P., Broom D.M. (2009) Measuring zoo animal welfare: Theory and practice. *Zoo Biology* 28(6): 531–544. doi:10.1002/zoo.20276

Jennings M., Prescott M.J., Members of the Joint Working Group on Refinement (Primates) (2009) *Refinements in husbandry, care and common procedures for non-human primates:* Ninth report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. *Laboratory Animals* 43(1): 1–47. doi:10.1258/la.2008.007143

Justice W.S.M., O'Brien M.F., Szyszka O., Shotton J., Gilmour J.E.M., Riordan P., Wolfensohn S. (2017) Adaptation of the animal welfare assessment grid (AWAG) for monitoring animal welfare in zoological collections. *Veterinary Record* 181(6): 143–143. doi:10.1136/vr.104309

Lawrence A.B., Vigors B., Sandøe P. (2019) What is so positive about positive animal welfare?—A critical review of the literature. *Animals* 9(10): 783. doi:10.3390/ani9100783

Line S., Markowitz H., Morgan K., Strong S. (1991) Effects of cage size and environmental enrichment on behavioral and physiological responses of rhesus macaques to the stress of daily events. In: Novak M.A., Petto A.J. (eds.). *Through the Looking Glass: Issues of Psychological Well-Being in Captive Nonhuman Primates*. Washington, DC: American Psychological Association, 160–179.

Lutz C.K., Novak M.A. (2005) Environmental enrichment for nonhuman primates: Theory and application. *ILAR Journal* 46(2): 178–191. doi:10.1093/ilar.46.2.178

Manteca Vilanova X. (2020) *The fundamentals of animal welfare assessments*. EAZA Annual Conference 2020 Online.

Meagher R.K. (2009) Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science* 119(1–2): 1–14. doi:10.1016/j.applanim.2009.02.026

Melfi V. (2005) The appliance of science to zoo-housed primates. *Applied Animal Behaviour Science* 90(2): 97–106. doi:10.1016/j.applanim.2004.08.017

Mellen J., MacPhee M.S. (2001) Philosophy of environmental enrichment: Past, present, and future. *Zoo Biology* 20(3): 211–226. doi:10.1002/zoo.1021

Mellor D.J. (2016) Updating animal welfare thinking: Moving beyond the "Five Freedoms" towards "A Life Worth Living." *Animals* 6(3): 21. doi:10.3390/ani6030021

Mellor D.J. (2017) Operational details of the Five Domains Model and its key applications to the assessment and management of animal welfare. *Animals* 7(8): 60. doi:10.3390/ani7080060

Mellor D.J., Beausoleil N.J. (2015) Extending the "Five Domains" model for animal welfare assessment to incorporate positive welfare states. *Animal Welfare* 24(3): 241–253. doi:10.7120/09627286.24.3.241

Mellor D.J., Beausoleil N.J., Littlewood K.E., McLean A.N., McGreevy P.D., Jones B., Wilkins C. (2020) The 2020 Five Domains Model: Including human–animal interactions in assessments of animal welfare. *Animals* 10(10): 1870. doi:10.3390/ani10101870

Mellor D.J., Reid C.S.W. (1994) Concepts of animal well-being and predicting the impact of procedures on experimental animals. *Improving the Well-being of Animals in the Research Environment* 3–18.

Mitchell G., Tromborg C.T., Kaufman J., Bargabus S., Simoni R., Geissler V. (1992) More on the 'influence' of zoo visitors on the behaviour of captive primates. *Applied Animal Behaviour Science* 35(2): 189–198.

Novak M.A., Kelly B.J., Bayne K., Meyer J.S. (2012) Behavioral disorders of nonhuman primates. In: Abee C.R., Mansfield K., Tardif S., Morris T. (eds.). *Nonhuman Primates in Biomedical Research, Second ed*. Boston, Massachusetts: Academic Press, 177–196. doi:10.1016/B978-0-12-381365-7.00007-8

Nuffield Council on Bioethics (ed.). (2005) The capacity of animals to experience pain, distress and suffering. In: *The Ethics of Research Involving Animals.* London, UK: Nuffield Council on Bioethics, 60–81.

Pomerantz O., Meiri S., Terkel J. (2013) Socio-ecological factors correlate with levels of stereotypic behavior in zoo-housed primates. *Behavioural Processes* 98, 85–91. doi:10.1016/j.beproc.2013.05.005

Rennie A.E., Buchanan-Smith H.M. (2006) Refinement of the use of non-human primates in scientific research. Part I: The influence of humans. *Animal Welfare* 15(3): 203–213. doi:10.1017/S096272860003044X

Richmond S.E., Wemelsfelder F., Beltrán de Heredia I., Ruiz R., Canali E., Dwyer C.M. (2017) Evaluation of animal-based indicators to be used in a welfare assessment protocol for sheep. *Frontiers in Veterinary Science* 4: 210. doi:10.3389/fvets.2017.00210

Roe K., McConney A., Mansfield C.F. (2014) The role of zoos in modern society—A comparison of zoos' reported priorities and what visitors believe they should be. *Anthrozoös* 27(4): 529–541. doi:10.2752/0892

Rose P., O'Brien M. (2020) Welfare assessment for captive Anseriformes: A guide for practitioners and animal keepers. *Animals* 10(7): 1132. doi:10.3390/ani10071132

RSPCA (2013) When coping is not enough: Promoting positive welfare states in animals. *RSPCA Australia Scientific Seminar 2013 Proceedings.* Canberra, Australia: RSPCA, 68.

Schmidt M. (2011) Locomotion and postural behavior. *Advances in Science and Research* 5(1): 23–29. doi:10.5194/asr-5-23-2010

Sherwen S.L., Hemsworth L.M., Beausoleil N.J., Embury A., Mellor D.J. (2018) An animal welfare risk assessment process for zoos. *Animals* 8(8): 130. doi:10.3390/ani8080130

Sherwen S.L., Hemsworth P.H. (2019) The visitor effect on zoo animals: Implications and opportunities for zoo animal welfare. *Animals* 9(6): 366. doi:10.3390/ani9060366

Velarde A., Geers R., European Cooperation in the Field of Scientific and Technical Research (Organization), Working Group 2: On Farm Monitoring of Welfare, Subworking Group: Pigs (2007) *On Farm Monitoring of Pig Welfare*. Wageningen, Netherlands: Wageningen Academic Publishers.

Vieira A., Battini M., Can E., Mattiello S., Stilwell G. (2018) Inter-observer reliability of animal-based welfare indicators included in the Animal Welfare Indicators welfare assessment protocol for dairy goats. *Animal* 12(9): 1942–1949. doi:10.1017/S1751731117003597

WAZA (2015) *Caring for wildlife: The World Zoo and Aquarium Animal Welfare Strategy.* Gland, Switzerland: World Association of Zoos and Aquariums Executive Office.

Webster J. (2005) *Animal Welfare: Limping Towards Eden: A Practical Approach to Redressing the Problem of Our Dominion Over the Animals.* Chichester, UK: Wiley-Blackwell.

Webster J. (1995) *Assessment of animal welfare: The Five Freedoms. In: Webster J., Animal Welfare: A Cool Eye towards Eden.* Oxford, UK: Blackwell Science, 10–14.

Weiss D., Hampshire V. (2015) Primate wellness exams. *Lab Animal* 44: 342–344. doi:10.1038/laban.835

Welfare Quality® (2009a) *Welfare Quality® Assessment Protocol for Pigs (Sows and Piglets, Growing and Finishing Pigs)*. Lelystad, Netherlands: Welfare Quality® Consortium.

Welfare Quality® (2009b*) Welfare Quality® Assessment Protocol for Poultry (Broilers, Laying Hens).* Lelystad, Netherlands: Welfare Quality® Consortium.

Welfare Quality® (2009c) *Welfare Quality® Assessment Protocol for Cattle.* Lelystad, Netherlands: Welfare Quality Consortium.

Welfare Quality® (2007) *Principles and criteria of Good Animal Welfare.* Lelystad, Netherlands: Welfare Quality Consortium.

Wells D.L. (2005) A note on the influence of visitors on the behaviour and welfare of zoo-housed gorillas. *Applied Animal Behaviour Science* 93(1–2): 13–17. doi:10.1016/j.applanim.2005.06.019

Whitham J.C., Wielebnowski N. (2009) Animal-based welfare monitoring: Using keeper ratings as an assessment tool. *Zoo Biology* 28(6): 545–560. doi:10.1002/zoo.20281

Whitham J.C., Wielebnowski N. (2013) New directions for zoo animal welfare science. *Applied Animal Behaviour Science* 147(3–4): 247–260. doi:10.1016/j.applanim.2013.02.004

Wildlife Reserves Singapore (n.d.) *Welfare Assessment*. Wildlife Reserves Singapore. Available online at https://www.eaza.net/assets/Uploads/Areas-of-Activity/Animal-welfare/Docs/WRS-Animal-Welfare-Assessment.pdf (accessed 5 July 2021).

Wolfensohn S., Sharpe S., Hall I., Lawrence S., Kitchen S., Dennis M. (2015) Refinement of welfare through development of a quantitative system for assessment of lifetime experience. *Animal Welfare* 24(2): 139–149. doi:10.7120/09627286.24.2.139

Wolfensohn S., Shotton J., Bowley H., Davies S., Thompson S., Justice W.S.M. (2018) Assessment of welfare in zoo animals: Towards optimum quality of life. *Animals* 8(7): 110. doi:10.3390/ani8070110

Wongpakaran N., Wongpakaran T., Wedding D., Gwet K.L. (2013) A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology* 13: 61. doi:10.1186/1471-2288-13-61

Wood W. (1998) Interactions among environmental enrichment, viewing crowds, and zoo chimpanzees (*Pan troglodytes*). *Zoo Biology* 17(3): 211–230. doi:10.1002/(SICI)1098-2361(1998)17:3<211::AID-ZOO5>3.0.CO;2-C

Yeates J.W., Main D.C.J. (2008) Assessment of positive welfare: A review. *The Veterinary Journal* 175(3): 293–300. doi:10.1016/j.tvjl.2007.05.009

Yon L., Williams E., Harvey N.D., Asher L. (2019) Development of a behavioural welfare assessment tool for routine use with captive elephants. *PLoS ONE* 14(2): e0210783. doi:10.1371/journal.pone.0210783

Zec S., Soriani N., Comoretto R., Baldi I. (2017) High agreement and high prevalence: The paradox of Cohen's kappa. *The Open Nursing Journal* 11(1): 211–218. doi:10.2174/1874434601711010211