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

EAZA

JZAR Research article

OPEN ACCESS

Research article

Feasibility and validity of the Animal Welfare Assessment Grid to monitor the welfare of zoo-housed gorillas *Gorilla gorilla gorilla*

Stijn P. Brouwers and Marie José H. M. Duchateau

Animal Behaviour and Cognition, Utrecht University, Utrecht, the Netherlands

Correspondence: Stijn P. Brouwers, email; Stijn.Brouwers@agroscope.admin.ch

Keywords: Animal Welfare Assessment Grid; keeper assessments; welfare monitoring; western lowland gorilla; zoo animal welfare

Article history:
Received: 05 Feb 2021
Accepted: 09 Sept 2021
Published online: 31 Oct 2021

Abstract

Zoos need to monitor their animals in order to evaluate to what extent animal welfare policies result in adequate welfare. Since it is usually not feasible for zoos to structurally measure corticosteroid concentrations or conduct extensive behavioural observations, zoos often rely on their caretakers to assess animal welfare using surveys. The Animal Welfare Assessment Grid (AWAG) allows zoos to quantify and visualise animal welfare based on keeper ratings. This tool has previously been used to monitor the welfare of zoo-housed animals, but it has not yet been used in practice by zookeepers. Therefore, the welfare of two groups of western lowland gorillas *Gorilla gorilla gorilla* was monitored daily for three months by caretakers using the AWAG to assess its usability and reliability. Behavioural observations were conducted simultaneously to validate keeper ratings of animal-based welfare indicators. This study demonstrated that the AWAG can be used to get a good indication of the welfare of an individual or group and to identify potential welfare issues. Welfare appeared to be relatively stable in the long term, which indicates that it is not necessary to perform daily welfare audits. Keepers' assessments captured more subtle changes in welfare compared to assessments made retrospectively by researchers in previous studies. Inter-rater reliability was good, but caretakers' scores did not always correspond with data from behavioural observations. Extra training, regular staff meetings and longer observation times will most likely increase the degree of detail of keeper ratings.

Introduction

In zoological institutions, ensuring animal welfare is primarily important for the well-being of the animals, but it also contributes to the conservation, education and research purposes of zoos (IUDZG/CBSG [IUCN/SSC] 1993; Powell and Watters 2017). Animal welfare can be defined as an animal's combined physical, mental and emotional state as perceived by the animal itself over a period of time (Harley and Clark 2019; AZA 2020). It can range from negative/bad welfare – impairing the animal – to positive/good welfare (Ohl and van der Staay 2012). Several zoo and aquarium associations worldwide already require their members to have a clearly documented animal welfare policy (e.g., AZA 2020; BIAZA 2020). In order for zoos to evaluate to what extent their welfare policy actually results in good welfare, the well-being of the animals in their care needs to be monitored.

Traditional methods of quantifying animal welfare, like the measurement of corticosteroid concentrations or conducting extensive behavioural observations, are often not practical for zoos due to limited time, resources and expertise (Hill and Broom 2009). Therefore, zoos usually rely on keepers to monitor animal welfare (Binding et al. 2020). Keepers have been identified as valuable proxies of zoo animal welfare, since they generally have years of experience with particular species and individuals (Whitham and Wielebnowski 2013; Marchant-Forde 2015). Their assessment of animal welfare can be divided in resource- and animal-based indicators. Resource-based welfare indicators consider the input that is provided to the animal, like housing and diet, while animal-based indicators cover the output, which is the animal's physical state and behaviour (Whay 2007). Although proper resources are essential to facilitate well-being, they do not guarantee that animals actually experience good welfare (Barber 2009; Ward et al. 2018). Therefore, welfare surveys used by keepers should ideally consist of both resource- and animal-based welfare indicators. To provide a complete assessment of welfare, these surveys need to contain species-specific welfare indicators.

A variety of welfare assessments based on keeper ratings has been developed specifically for zoo animals (Sherwen et al. 2018). Kagan et al. (2015) have elaborated a welfare assessment checklist consisting of environmental, physical and psychological welfare indicators. However, this method does not take into account any species-specific needs, even though zoos house a wide range of taxa that each have their own specific needs for resources and behaviour (Wolfensohn et al. 2018). WelfareTrak does use species-specific surveys to assess and monitor the welfare of zoo animals (Whitham and Wielebnowski 2009; 2013). Nonetheless, there are currently surveys available for only a limited number of species, making this tool not yet suitable for zoos to implement on a large scale. The '24/7' approach to zoo animal welfare (Brando and Buchanan-Smith 2018) can theoretically be used to evaluate the welfare of all zoo-housed animals, given that sufficient biological information is available on the species. However, this method does not use numerical scores that facilitate quantitative monitoring of welfare.

The Animal Welfare Assessment Grid (AWAG) is a welfare assessment tool that is ready to use, considers species-specific needs and uses numerical welfare scores. This method was originally developed for laboratory primates (Honess and Wolfensohn 2010; Wolfensohn et al. 2015), but it has been adapted for use on zoo-housed primates and birds (Justice et al. 2017). The assessment is intended for use by zoo staff and provides a method to quantify welfare based on a combination of general and species-specific welfare indicators. The survey consists of a physical, psychological, environmental and medical procedural class of welfare indicators, and includes both resource- and animal-based indicators. All 22 welfare indicators are scored on a 10-point scale, of which each value is defined to minimise inter-rater bias. The AWAG's strongest feature is the ability to visualise animal welfare data (Wolfensohn et al. 2018). For any given timeframe, the averages of the four parameter classes (physical, psychological, environmental and procedural) can be plotted as a radar chart to form a two-dimensional polygon which represents the impact of each category on an animal's welfare. The Cumulative Welfare Assessment Score (CWAS) is equal to the surface area of this radar chart and is thus not just the average of the four parameter classes. The CWAS increases exponentially instead of linearly when multiple parameter classes are compromised, indicating a potential large welfare issue. While the radar chart can be used to capture long-term trends in animal welfare, the CWAS can be plotted over time to identify short-term events that impact well-being. This makes the AWAG one of the most advanced methods currently available to assess and visualise animal welfare data for zoos.

However, the AWAG has not been used in practice by animal caretakers, since researchers completed welfare surveys retrospectively based on lifetime records or keeper reports in previous studies (Wolfensohn et al. 2015; Justice et al. 2017). Therefore, the aim of this study was to assess the usability and reliability of the AWAG as animal welfare assessment tool for zoos by letting caretakers monitor the welfare of two groups of captive western lowland gorillas *Gorilla gorilla gorilla*. Additionally, the study aimed to validate keeper scores by correlating ratings of animal-based welfare indicators with data from concurrent behavioural observations made by a researcher.

Methods

Subjects of this study were eight adult western lowland gorillas housed in two separate groups in Safaripark Beekse Bergen, Hilvarenbeek, the Netherlands. One group was a family group and consisted of one adult male (silverback), three adult females and one infant female. The other group was an all-male group and consisted of four adult males (all silverbacks), who were (half) brothers. One of these males (MB) suffered from progressive retinal degeneration and was euthanised during the study period (this decision was not based on results of the current study). Both groups were housed in similar enclosures, each consisting of an indoor and outdoor exhibit area within the public's view and four interconnected holding areas out of the public's view. The indoor enclosures contained extensive climbing structures and visual barriers. The outdoor enclosures were surrounded by a wet moat and contained climbing structures, rock outcroppings, small shelters and multiple trees surrounded by electric fences to prevent the gorillas from climbing in them. The outdoor enclosure of the family group had less vegetation than that of the all-male group. The family group was housed together with three female black-crested mangabeys *Lophocebus aterrimus* and the all-male group was housed together with six male eastern black-and-white colobus *Colobus guereza*. The gorillas were fed multiple times a day and water was available ad libitum. The gorillas were on public display every day between approximately 1000 and 1630 h. However, due to the restrictions regarding the COVID-19 pandemic, the park was closed for visitors between 14 March and 15 May 2020.

Welfare monitoring

The welfare of the adult gorillas was monitored from 5 March to 5 June 2020 using the AWAG scoring system of Justice et al. (2017), with species-specific adjustments based on the EAZA best practice guidelines for western lowland gorillas (Abelló et al. 2017). Welfare was assessed individually at the end of each day by the caretakers that worked with the gorillas the most that day. However, due to a lack of time, the assessments could not be carried out on seven days for the family group and on 11 days for the all-male group. In total, there were 332 individual assessments completed in the family group and 289 individual assessments in the all-male group. Eight different zookeepers had completed the assessments.

The welfare assessment consisted of 22 welfare indicators divided into four parameter classes; physical, psychological, environmental and procedural (described below and in the supplements). Each parameter was assigned a factor score between 1 and 10, with 1 being the best possible state relative to a healthy individual of the same sex and age and 10 the worst state. Extensive definitions were provided for each factor score to ensure that the scales were as balanced as possible and to minimise inter-rater bias (Supplementary Tables 1–4). Since the length of each arm of the radar chart represents the impact of that parameter class on the animal's welfare, low factor scores represented optimal welfare conditions. Based on pilot data and initial keeper feedback, it was decided to assign expected values to the welfare indicators for each individual. Expected values were proposed by the researcher in consultation with the keepers based on visual inspections of the animals and their enclosures. When scores deviated from their expected value, additional clarification was provided by the keeper through a written comment.

Physical parameters

For the physical parameter class, five different animal-based indicators were assessed: general condition (weight, body condition score and coat condition), clinical assessment (including signs such as injury, alopecia and vomiting), faecal consistency,

Table 1. Ethogram of state and event behaviours (adapted from van den Berg et al. 2018).

Solitary	Behaviour	Definition	Type	Receiver
	Forage (for regular food)	Searching for, handling or consuming vegetables provided by keepers or vegetation growing in outside enclosure	State	No
	Use foraging enrichments	Handling tubes/barrels filled with food or consuming food derived from these enrichments, or searching for, handling or consuming branches provided by keepers	State	No
	Inactive	Sitting, lying down, hanging, or sleeping; includes nest building	State	No
	Move	Walking or running from one location to another, arboreal or terrestrial	State	No
	Object play	Handling, examining or manipulating non-food object(s)	State	No
	Self-groom	Manipulation of own fur/skin with hand, foot or mouth; does not include scratching	State	No
	Self-play	Toying with own body or doing acrobatics; not directed at other animals; can include summersaults, jumping, playful running, thigh slap, chest beat play and pirouetting	State	No
	Out of sight	The focal animal is not visible to the observer	State	No
Affiliation	Allo-groom	Oral or manual manipulation of hair or skin of another individual; already implies contact-sitting	State	Yes
	Social play	Non-aggressive active interaction between two or more individuals with behaviours such as wrestling or running after each other	State	Yes
	Chest beat play	Striking or hitting own body/chest with hands (includes ‘incomplete’ chest beat) during or 30 sec before/after play bout; does not always have a clear receiver	Event	Yes/No
Contact aggression	Bite	Using mouth/teeth to bite another individual in an aggressive way	Event	Yes
	Hit	Forcefully slapping another individual using hands or feet	Event	Yes
	Pull	Forcefully pulling or grabbing the fur or body part of another individual using hands or feet	Event	Yes
	Push	Forcefully shoving another individual using hands or feet	Event	Yes
Non-contact aggression	Chest beat	Striking or hitting own body/chest with cupped/flat hands; does not always have a clear receiver	Event	Yes/No
	Display	Assuming a dominant body posture (quadrupedal stance or standing on hind legs with/without chest beating), open mouth exposing teeth/canines, strutting and/or aggressively striking object(s); does not always have a clear receiver	Event	Yes/No
Stress behaviours	Displacement	Actively approaching another individual in a non-neutral way causing the other individual to move away or withdraw within 5 sec	Event	Yes
	Chase	Pursuing and running after another individual at high speed (running; not social play)	Event	Yes
	Scratch	Fast movement of foot/hand against own body	Event	No
	Yawn	Opening mouth to yawn	Event	No
	Nose wipe	(Repetitive) brushing of the nose with hand/foot	Event	No
Abnormal behaviours	Coprophagy	Eating faeces	Event	No
	Regurgitation and re-ingestion	Vomiting and eating vomit again (ingestion; can occur inside the mouth as well)	Event	No
Stereotypy	Stereotypical behaviours	Sucking on digit, lip, others or objects, rocking and pacing, hair plucking, wound picking, bizarre body posturing (e.g., holding head) and self-clasping	Event	No

activity level (including assessment of mobility) and food/water intake (including perceived hunger and thirst) (Supplementary Table 1). General condition was mainly scored based on visual inspections, since the animals were weighed not more than once per year. Apart from minor modifications to factor score definitions, the welfare indicators within this parameter class did not differ from the ones scored by Justice et al. (2017).

Psychological parameters

Psychological well-being was assessed using six, predominantly animal-based, welfare indicators: abnormal behaviours, aggression, (foraging) enrichment provision and use, reaction to routine events (moving animals in/out night enclosures etc.), animal training and anticipatory behaviour (Supplementary Table 2). Anticipatory behaviour is a new indicator, added because of

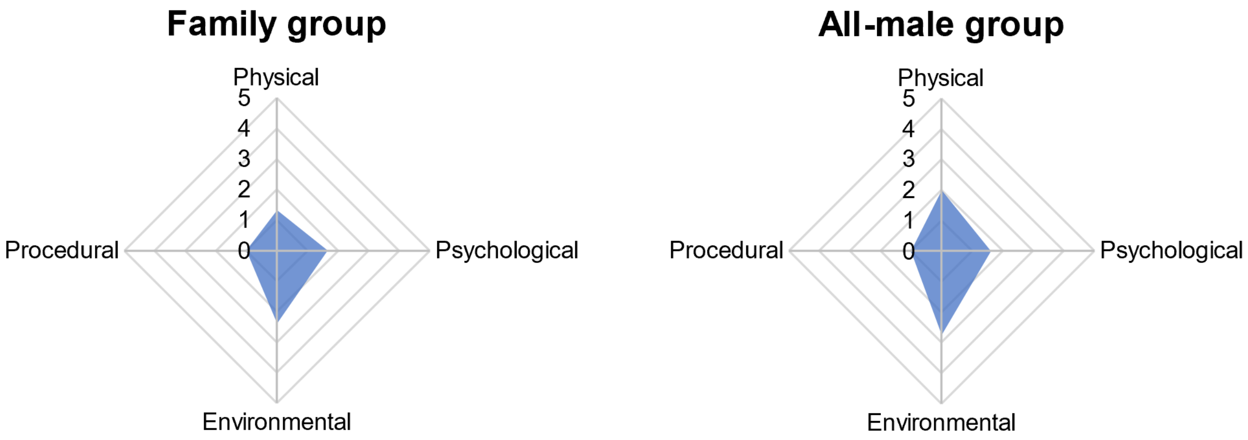

Figure 1. Animal welfare assessment grids of the family group (n=4) and the all-male group (n=4). Radar chart represents the average scores for the physical, psychological, environmental and procedural parameter classes over the entire study period on a scale from 1 to 10, with 1 being the best possible state and 10 the worst. In the figure, the axes are adjusted based on the range of the average parameter class scores. The area under the AWAG equates to the CWAS for the whole study period.

its potentially large value to animal welfare assessment (Watters 2014; Krebs et al. 2017; Clegg et al. 2018). This behaviour is common in zoo animals since daily husbandry events are generally scheduled or preceded by obvious cues. ‘Response to catching event’ was omitted in the current study, because adult gorillas in captivity are virtually never actively caught without sedation.

Environmental parameters

The environmental parameter class included eight resource-based indicators: housing, group size, furnishing/enclosure design, nutrition, space, introductions, contingent events and (perceived) control (Supplementary Table 3). ‘(Perceived) control’ was added as a welfare indicator, because it is becoming increasingly clear that choice and control over the environment can be critical to welfare (Leotti et al. 2010). The assessment of this indicator was based on the access animals had to different parts of their enclosure (Ross 2006; Kurtycz et al. 2014), the complexity of their physical environment (Ross et al. 2011), the temporal complexity of their day (i.e., predictability of scheduled events), whether they had the option to remain out of sight of visitors and the nature of their relationships with caretakers (Carlstead 2009).

Procedural parameters

Within the procedural parameter class, the welfare indicators were primarily resource-based; sedation, veterinary procedure and change in daily routine (Supplementary Table 4). However, these risk factors were also partially animal-based indicators since score definitions included the recovery of the animal after a medical procedure. ‘Restraint’ was not included during our study as no gorillas were caught without sedation.

Behavioural observations

To validate the accuracy of keeper ratings, behavioural observations of focal individual gorillas were performed during the study by one researcher, who was not affiliated to the zoo. For intra-observer reliability, the researcher coded the same 10-min video of the all-male group before and after the study period using the ethogram described below. This resulted in an excellent agreement (Cohen’s Kappa, κ=0.869; Cicchetti 1994).

For each group, behavioural observations were conducted weekly between 1000 and 1600 h, using a randomised observation schedule. Observations were not carried out during feeding presentations. Due to the restrictions regarding the COVID-19 pandemic, no observations could be conducted from 20 March to 6 May. Individuals were observed from public viewing areas in 15-min focal watches (Martin and Bateson 2007), using a predefined ethogram (Table 1) in BORIS (Friard and Gamba 2016). Foraging behaviour was divided into foraging for regular food and using foraging enrichments, since ‘use of foraging enrichments’ is a separate indicator in the AWAG (Supplementary Table 2). When the focal animal went out of sight, the observation was paused and continued when the individual was back in sight. If the animal remained out of sight for longer than two minutes, the observation was terminated and the remaining time was added to the next observation or carried out as a separate observation.

For the family group, 120 15-min focal watches were collected, which yielded 7.5 hours of observational data per individual. The infant female was not observed since her welfare was not monitored. For the all-male group, 113 15-min focal watches were collected, which yielded 8.5 hours of observational data per individual. The euthanised male (MB) was observed for a total of 2.75 hours. To check for inter-rater reliability, the researcher completed the welfare assessment separately from a gorilla caretaker on observation days, resulting in 24 individual assessments at each group.

Data analyses

For each individual, average daily scores were calculated for all 22 welfare indicators. The daily averages of the parameter class scores were subsequently plotted as a radar chart to generate

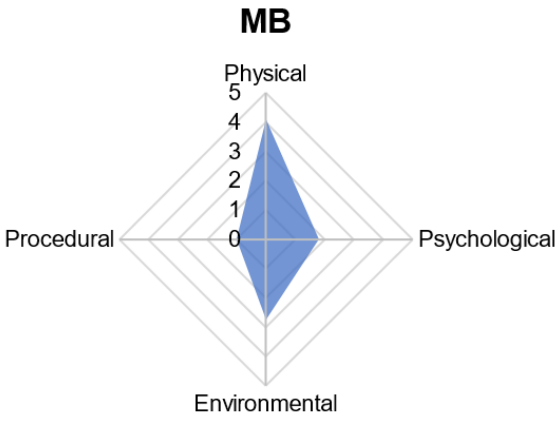

Figure 2. Individual animal welfare assessment grid of MB, the male in the all-male group who suffered from progressive retinal degeneration and was euthanised during the study period.

individual AWAGs for each day of the study period. Based on the average parameter class scores of the entire study period, individual and group AWAGs were created that represented the average welfare state over the total study period. The limits of the axes were adjusted based on the range of the average parameter class scores. The area under the daily AWAG equated to the CWAS for that day. For each individual, the CWAS was plotted over the course of the study to visualise variability in welfare over time. Days on which the assessment was not completed were left blank.

Activity budgets were constructed for each focal individual using state behaviours to describe the proportion of time spent on different behaviours. Additionally, rates of event behaviours were calculated as frequency per hour. The accuracy of keeper assessments was investigated by correlating ratings of animal-based welfare indicators with data from the behavioural observations. On observation days, the researcher also completed the welfare assessment separately from the gorilla keeper. Inter-rater agreement was examined by determining the percentage of scores that differed between the keeper and the researcher.

Data were statistically analysed using R Statistics Software version 4.0.0. To estimate inter-rater reliability, an intraclass correlation (ICC; assuming two-way random single measures) was used to compare ratings of keepers and ratings of the researcher (Bartko 1966). For all statistical tests, the threshold of significance was set at P<0.05.

Results

Welfare monitoring

Average indicator scores, parameter class scores and cumulative welfare scores are given in Supplementary Table 5. Figure 1 shows the AWAGs of both groups (average parameter class scores of the entire study), which were largely comparable in shape: low (i.e., good) average scores for the physical and procedural parameter classes, and somewhat higher (i.e., suboptimal) average scores for the psychological and environmental parameter classes. The higher average physical score of the all-male group is caused by MB, who scored considerably higher for physical indicators than his group members due to his disease. The all-male group also had a slightly higher score for the environmental parameter class than the family group due to a higher average score for Group size, since all-male groups are considered to be less stable than family groups (Stoinski et al. 2004).

The individual AWAG of the (partially) blind male MB had a considerably higher physical score compared to both group AWAGs (score MB: 4.10; mean family group: 1.32; mean all-male group: 1.99; Figure 2). This was not only caused by his visual impairment (scored under Clinical assessment; score 7.04), but also by a related loss in body weight (scored under General condition; score 7.04). Additionally, he scored higher for Activity level compared to the other gorillas (score MB: 4.17; mean: 1.48). Although his psychological score, indicative of mental health, was also slightly higher than the average of his group, this difference was not substantial (score MB: 1.82; mean all-male group: 1.63).

The cumulative welfare assessment scores of the gorillas in the

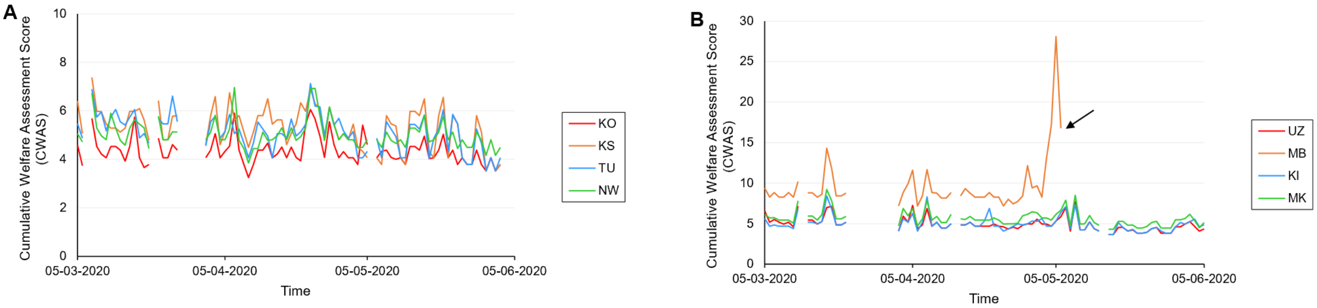

Figure 3. Daily cumulative welfare assessment scores over time for (A) each of the four adult individuals in the family group and (B) the four individuals in the all-male group. The black arrow indicates the euthanasia of the male MB. Days on which the assessment was not completed are left blank.

family group all followed the same pattern (Figure 3A). The CWAS of the silverback (KO) was consistently slightly lower than that of the females in his group, mainly due to a better average physical score (score KO: 1.04; mean females: 1.41; Supplementary Table 5). Despite day-to-day fluctuations, the CWAS of the individuals in the family group remained roughly within certain boundaries (approximately 4 to 7) throughout the study period and appeared to be relatively stable in the long term. For the all-male group, the daily cumulative welfare scores of the three healthy males were largely similar (Figure 3B). The CWAS of MB was structurally higher than that of the other males and increased considerably during the days preceding his death. The periods of larger spikes were mainly caused by higher scores for Hierarchy upset/dispute, aggression/bullying on those days.

Validation

Of the 968 welfare indicators scored by both keeper and researcher, 871 indicators had a matching score (89.98% agreement). The correlation between the scores given by keepers and the researcher was highly significant (intraclass correlation, ICC=0.92, P<0.001; Figure 4).

To validate the accuracy of caretaker surveys, the average keeper ratings of four different animal-based welfare indicators were correlated to relevant behavioural data recorded during observations. There was a significant correlation between the frequencies of aggressive behaviours observed by the researcher and average keeper ratings of Hierarchy upset/dispute, aggression/bullying (Pearson correlation, r=0.87, P=0.005, n=8; Figure 5A). There was no significant correlation between frequencies of abnormal behaviours and keeper ratings of Abnormal behaviours (Kendall rank correlation, r=0.40, P=0.190, n=8; Figure 5B). However, only one individual (TU) regularly performed abnormal behaviours according to the keepers (score 3.52; see also Supplementary Table 5), and this individual also performed the most abnormal behaviours during observations. There was no significant correlation between observed proportions of time spent on inactivity and keeper ratings of Activity level (Kendall rank correlation, r=0.46, P=0.123, n=8; Figure 5C). Caretakers assessed the activity level of all individuals as normal, except for MB, which was assessed as quite inactive (score 4.17; see also Supplementary Table 5). However, the researcher found no big difference between MB's activity level and that of the other gorillas. Lastly, there was no significant correlation between proportions of time spent on using foraging enrichments and keeper ratings of Use of (foraging) enrichments (Spearman rank correlation, r=−0.06, P=0.887, n=8; Figure 5D). For this indicator, caretakers scored between 2.5–3.0, indicating that multiple types of enrichment were available, but not always used. The researcher found that the average time spent using foraging enrichment differed significantly between the family group (μ: 8.4±5.1%) and the all-male group (μ: 23.4±3.3%) (Wilcoxon rank-sum test, W=16, P=0.029).

Discussion

The aim of this study was to determine whether the Animal Welfare Assessment Grid (AWAG) provides a practical, reliable and valid method for zoos to monitor the welfare of their animals. Therefore, the individual welfare of eight adult western lowland gorillas was monitored daily using the AWAG for approximately 3 months by zookeepers. Since the use of the AWAG by animal care staff had not been validated in previous studies (Wolfensohn et al. 2015; Justice et al. 2017), behavioural observations were conducted to assess the accuracy of keeper ratings of animal-based welfare indicators.

The current study has demonstrated that the AWAG can provide zoos with a practical method to get an indication of the welfare of the animals in their care, by visualising the average states of physical, psychological, environmental and procedural aspects of welfare in one figure. For example, the AWAG of a gorilla suffering from progressive retinal degeneration was for the physical parameter class markedly different than the AWAG of the other gorillas, which indicates that long-term impairments of physical welfare are well captured by the AWAG. The AWAG also suggested that other aspects of his welfare, like mental health, were not compromised. The AWAG of the all-male group did not differ markedly from that of the family group, indicating that having only males in a gorilla group does not necessarily cause welfare problems (Stoinski et al. 2004). The AWAG can potentially be used by zoos to identify possible welfare issues, compare the welfare of individuals or groups, and it facilitate informed decision making.

Where the AWAG visualises the average welfare over a given period, zoos can also monitor temporal variation in welfare by plotting the Cumulative Welfare Assessment Score (CWAS) over time. This can be done to identify events that impacted welfare or to evaluate the effectiveness of efforts to improve welfare. For example, the welfare of the males in the all-male group increased slightly after the euthanasia of their group member, since captive gorilla groups of two to three adult males are reported to be more stable than larger groups (Stoinski et al. 2004). The difference between the CWAS and other average welfare scores (e.g., Harley and Clark 2019) becomes apparent from the days preceding the euthanasia of MB. Although this gorilla's eyesight was severely reduced, his behaviour was initially not markedly different from that of the other males in his group. Therefore, only physical welfare indicators were assessed as below par. When he started to

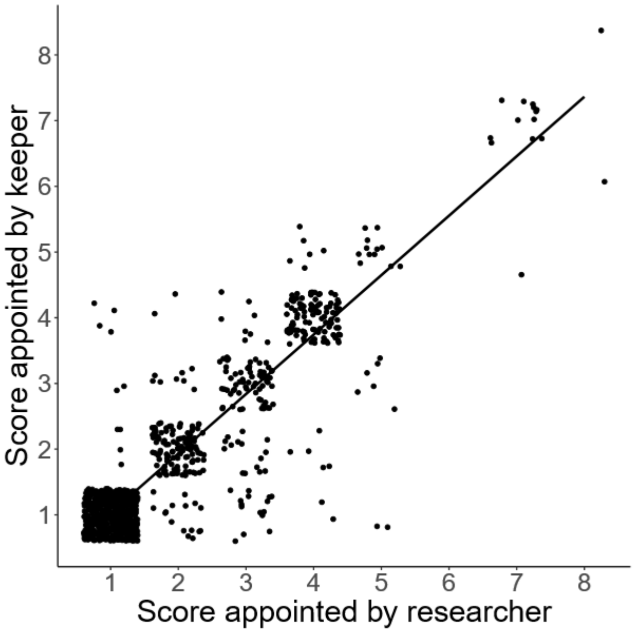

Figure 4. Inter-rater correlation between the individual welfare indicator scores appointed by keepers and scores appointed by the researcher (intraclass correlation, ICC=0.92, P<0.001). Dots are jittered to avoid overlap of data points.

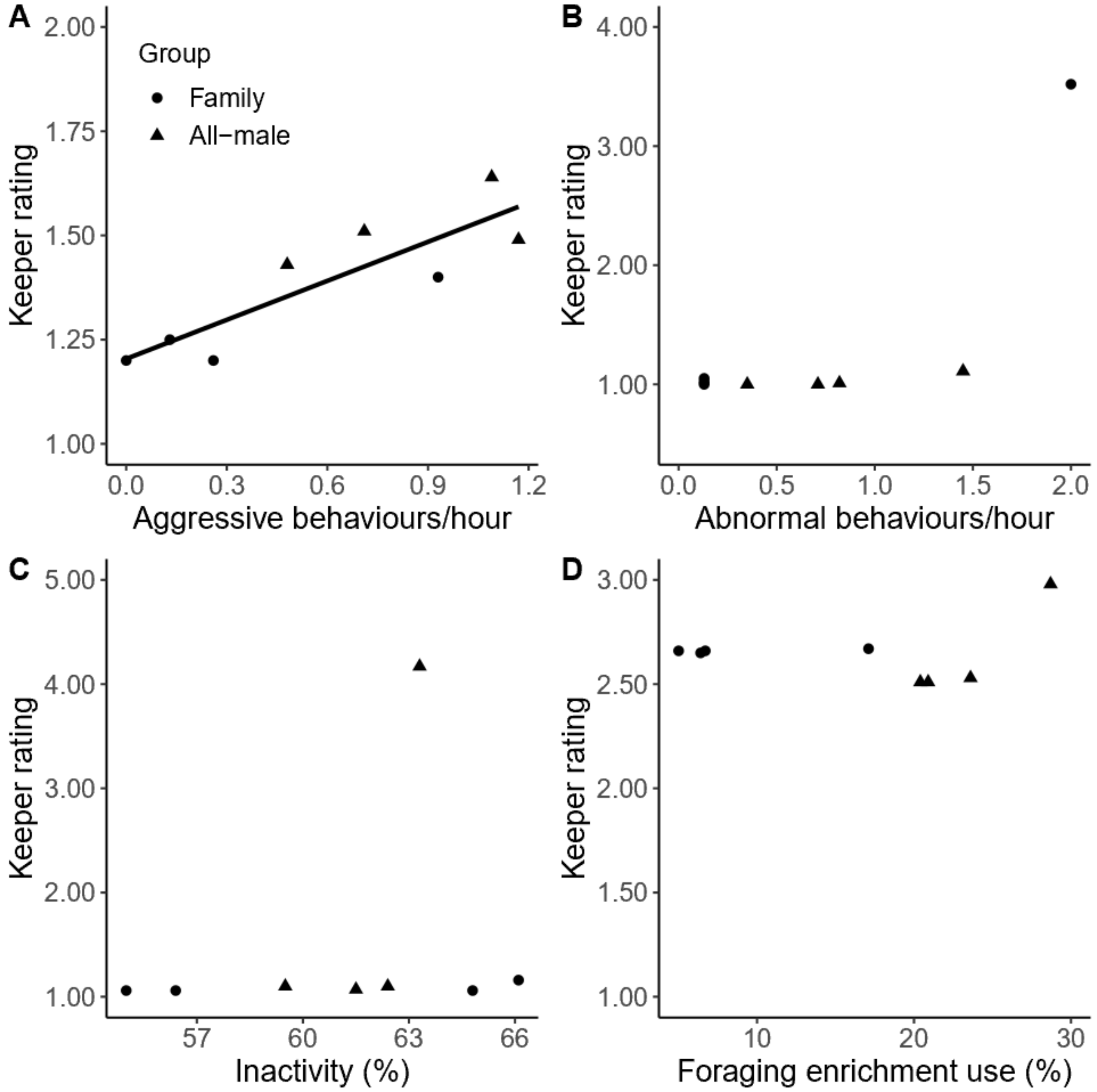

Figure 5. Correlation between (A) amounts of aggressive behaviours conducted per hour and keeper ratings of Hierarchy upset/dispute, aggression/bullying (Pearson correlation, r=0.87, P=0.005), (B) amounts of abnormal behaviours conducted per hour and keeper ratings of Abnormal behaviours (Kendall rank correlation, r=0.40, P=0.190), (C) percentages of time spent being inactive and keeper ratings of Activity level (Kendall rank correlation, r=0.46, P=0.123) and (D) percentages of time spent using foraging enrichments and keeper ratings of Use of (foraging) enrichments (Spearman rank correlation, r=−0.06, P=0.887). Dots indicate members of the family group, triangles indicate members of the all-male group.

perform abnormal behaviours in the days preceding his death, he also scored worse for psychological indicators. This combination of two parameter classes being compromised caused his CWAS to increase exponentially, since the CWAS is based on the surface area of the AWAG. In contrast, his last daily welfare score would be five times lower if the cumulative score was merely the average of the four parameter class scores, suggesting a much less profound welfare issue. Furthermore, the CWAS prevents animals who are not in good health, but still able to experience positive welfare (Mason and Mendl 1993), from automatically receiving poor welfare scores.

In comparison with previous studies (Wolfensohn et al. 2015; Justice et al. 2017), CWAS fluctuated more from day-to-day. The researchers in previous studies completed welfare assessments retrospectively based on lifetime records of animals or daily keeper reports. However, it is likely that only marked events and large impacts on welfare are included in written reports, while smaller changes may be overlooked in this way, resulting in less daily variation. The current study demonstrates that even subtle variations in welfare over time are captured by the AWAG when keepers complete the welfare assessment themselves. Moreover, it has previously been suggested that caretakers are the most suitable persons to evaluate animal welfare, because they know the preferences, temperaments, behaviour and routines of the animals in their care better than anyone (Whitham and Wielebnowski 2009). Additionally, it is most time- and cost-effective to assign keepers to complete welfare assessments.

The welfare of the gorillas in the current study was monitored daily. Their daily CWAS remained generally within a certain range throughout the study period and deviations from the baseline usually restored within several days. This suggests that their welfare was relatively stable in the long-term. If only one assessment per week is used in the data analysis, the average CWAS would differ 0.42 at most. This indicates that, at least for western lowland gorillas, it is not necessary to perform daily welfare assessments. If welfare audits are completed at a lower interval, zoos can use their limited time and resources to monitor the welfare of more animals. However, if welfare assessments are conducted at a lower intensity, variations across day and night, weekdays, weekends and seasons should still be taken into account, and audits should reflect the entire interval between assessments (Brando and Buchanan-Smith 2018; Sherwen et al. 2018).

Welfare assessments based on observer ratings can only be valuable if they produce data that are both reliable and valid (Meagher 2009). In the current study, inter-rater agreement between ratings of keepers and ratings of the researcher was high. This indicates that zookeepers were able to reliably assess the welfare indicators included in the AWAG, which is in accordance with previous studies in which animal caretakers reliably rated physical and behavioural characteristics of zoo animals (Carlstead et al. 1999; Wielebnowski 1999; Dutton 2008; Tetley and O'Hara 2012; Webb et al. 2020). In this study, score sheets with expected values were used, which most likely increased inter-rater agreement. During the evaluation, some keepers indicated that the expected values influenced their own view and thus reduced objectivity.

From the comparison between the ratings of zookeepers and the behavioural observations it can be concluded that the caretakers' scores are valid for the assessment of aggressive behaviour. Zookeepers are generally well aware of aggression, since it is often accompanied by vocalisations and possibly injuries. However, small differences in behavioural welfare indicators that occurred on a low frequency or that are less indicative, were not rated differently in the AWAG by the caretakers. For instance, caretakers appointed equal scores to both gorilla groups for Use of (foraging) enrichments, even though the researcher observed a significant difference in the time spent using foraging enrichments between the two groups. However, it was regularly discussed that the all-male group used foraging enrichments more than the family group during staff meetings, indicating that keepers were in fact aware of this difference. It is possible that only behaviours that differed markedly from expected values were memorised, since the assessment was completed at the end of the day.

This is in contrast with previous studies that validated caretakers' assessments of behaviours indicative of welfare against behavioural data (Stevenson-Hinde et al. 1980; Rousing and Wemelsfelder 2006; Yon et al. 2019). Even though animal-based welfare indicators rely on the scorer's subjective judgement, they are considered to be more indicative of welfare than resource-based indicators, since they directly provide insight into how well an animal is able to cope with its environment (Whay 2007). However, keepers in the current study generally observed animals during specific parts of the day, for example right after feeding them, while behavioural observations were not conducted around feeding times. It would help to give caretakers more time to observe their animals, and preferably at different moments during the day to improve the accuracy of ratings of behavioural welfare indicators included in the AWAG. Nonetheless, the AWAG needs to be validated more thoroughly in zoo settings by cross-validating keeper scores with different measures of welfare, such as behavioural diversity (Miller et al. 2020), or more direct indicators, such as cortisol measurements (Heimbürge et al. 2019) or judgement bias tests (Burman et al. 2011; Baciadonna and McElligott 2015).

This study attempted to incorporate anticipatory behaviour into the AWAG, since this is a behavioural measure of positive affective states and thus reflects an animal's own perception of its welfare (Watters 2014). However, keepers pointed out that this indicator was not fully clear to them, and more research is needed to formulate practical and balanced factor score definitions. Other suggested indicators of positive affective states include affiliative behaviours, sleep, play, vocalisations and exploratory behaviours (Whitham and Wielebnowski 2013). Sleep and vocalisations are less practical for zookeepers to assess, since caretakers are usually not present at night and there can be large inter-species differences in levels of vocalisations (e.g., gorillas call less frequently compared to chimpanzees; Byrne 1982). However, affiliative behaviours, play and exploratory behaviours can potentially be integrated into the AWAG to increase the focus on positive affective states (Mellor and Beausoleil 2015). To prevent misinterpretation of behavioural welfare indicators, keepers who will complete welfare assessments should receive additional training and proper supervision (Vasseur et al. 2013; Sherwen et al. 2018). Scheduling regular staff discussions is likely to be a good additional method of minimising subjectivity, achieving constant agreement, as well as adding free choice profiling (FCP) to capture subtle changes in behaviour, attitude and posture as suggested by Wemelsfelder et al. (2001).

In this study, possible scores ranged from 1 to 10. Although this allowed keepers to assess the indicators in detail, it also increased the time keepers needed to complete the welfare audit. Furthermore, certain score definitions differed minimally from each other and inter-rater reliability of animal welfare assessment tools has been reported to increase when scoring systems are simplified (Channon et al. 2009). Therefore, a 7- or 5-point scale may be more suitable to assess the welfare indicators included in the AWAG. A software application that is currently in development (Wolfensohn 2020 personal communication) will most likely increase the usability of the AWAG even further and make it more convenient for zoos to process animal welfare data.

Lastly, certain welfare indicators included in the AWAG can be a pitfall to scorers. For example, Activity level was generally assessed to be 'normal' for most gorillas and their activity levels indeed corresponded to previously reported activity levels of western lowland gorillas in zoos (Sarfaty et al. 2012; Racevska and Hill 2017). However, they were considerably less active than wild-living gorillas (Watts 1988; Masi et al. 2009), as is often the case for gorillas in captivity due to the reduced need to forage and the lower concentration of fibres in their diet (Masi 2011). To minimise confusion, clear guidelines should be set that state whether criteria should be scored relative to what is considered 'normal' for the species in captivity or for wild-living conspecifics.

Conclusion

The AWAG is a promising practical welfare assessment tool that, if developed further, can allow zoos to obtain a reliable assessment of the welfare of the animals in their care and identify potential welfare issues, based on a combination of resource- and animal-based welfare indicators. It is advised that caretakers receive additional training and proper supervision, regular staff meetings should be scheduled, and caretakers should be able to observe their animals for longer periods of time. The AWAG should be cross validated even more thoroughly using other measures of welfare and guidelines should be provided for scoring. In addition, future research should investigate whether the AWAG can be applied to other taxonomic groups housed in zoos. If so, the AWAG can be a valuable tool for zoos to monitor the welfare of their animals and ultimately improve their welfare standards.

Acknowledgements

The authors would like to thank Safaripark Beekse Bergen, with special acknowledgement to Kris Jansen and the animal caretakers of the gorillas for their contributions to the study.

References

Abelló M.T., Rietkerk F., Bemment N. (2017) *EAZA Best Practice Guidelines for Gorillas*, 2nd ed.

Association of Zoos and Aquariums (2020) The Accreditation Standards and Related Policies. Available at www.aza.org (accessed October 31, 2020).

Baciadonna L., McElligott A.G. (2015) The use of judgement bias to assess welfare in farm livestock. *Animal Welfare* 24: 81–91.

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

Bartko J.J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19(1): 3–11.

Binding S., Farmer H., Krusin L., Cronin K. (2020) Status of animal welfare research in zoos and aquariums: Where are we, where to next? *Journal of Zoo and Aquarium Research* 8(3): 166–174.

Brando S., Buchanan-Smith H.M. (2018) The 24/7 approach to promoting optimal welfare for captive wild animals. *Behavioural Processes* 156: 83–95.

British and Irish Association of Zoos and Aquariums (2020) BIAZA Animal Welfare Policy. Available at www.biaza.org.uk (accessed November 2, 2020).

Burman O., McGowan R., Mendl M., Norling Y., Paul E., Rehn T., Keeling L. (2011) Using judgement bias to measure positive affective state in dogs. *Applied Animal Behaviour Science* 132(3-4): 160–168.

Byrne R.W. (1982) Primate vocalisations: Structural and functional approaches to understanding. *Behaviour* 80: 241–258.

Carlstead K., Mellen J., Kleiman D.G. (1999) Black rhinoceros (*Diceros bicornis*) in US zoos: I. Individual behavior profiles and their relationship to breeding success. *Zoo Biology* 18(1): 17–34.

Carlstead K. (2009) A comparative approach to the study of keeper-animal relationships in the zoo. *Zoo Biology* 28(6): 589–608.

Channon A.J., Walker A.M., Pfau T., Sheldon I.M., Wilson A.M. (2009) Variability of Manson and Leaver locomotion scores assigned to dairy cows by different observers. *Veterinary Record* 164(13): 388–392.

Cicchetti D.V. (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4): 284.

Clegg I.L., Rödel H.G., Boivin X., Delfour F. (2018) Looking forward to interacting with their caretakers: Dolphins' anticipatory behaviour indicates motivation to participate in specific events. *Applied Animal Behaviour Science* 202: 85–93.

Dutton D.M. (2008) Subjective assessment of chimpanzee (*Pan troglodytes*) personality: Reliability and stability of trait ratings. *Primates* 49(4): 253–259.

Friard O., Gamba M. (2016) BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution* 7(11): 1325–1330.

Harley J., Clark F.E. (2019) Animal Welfare Toolkit. London. BIAZA.

Heimbürge S., Kanitz E. Otten W. (2019) The use of hair cortisol for the assessment of stress in animals. *General and Comparative Endocrinology* 270: 10–17.

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

Honess P., Wolfensohn S. (2010) The extended welfare assessment grid: A matrix for the assessment of welfare and cumulative suffering in experimental animals. *Alternatives to Laboratory Animals* 38(3): 205–212.

IUDZG/CBSG (IUCN/SSC) (1993) The world zoo conservation strategy: The role of the zoos and aquaria of the world in global conservation.

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.

Kagan R., Carter S., Allard S. (2015) A universal animal welfare framework for zoos. *Journal of Applied Animal Welfare Science* 18(sup1): S1–S10.

Krebs B.L., Torres E., Chesney C., Kantoniemi Moon V., Watters J.V. (2017) Applying behavioral conditioning to identify anticipatory behaviors. *Journal of Applied Animal Welfare Science* 20(2): 155–175.

Kurtycz L.M., Wagner K.E., Ross S.R. (2014) The choice to access outdoor areas affects the behavior of great apes. *Journal of Applied Animal Welfare Science* 17(3): 185–197.

Leotti L.A., Lyengar S.S., Ochsner K.N. (2010) Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences* 14(10): 457–463.

Marchant-Forde J.N. (2015) The science of animal behavior and welfare: Challenges, opportunities, and global perspective. *Frontiers in Veterinary Science* 2: 16.

Martin P., Bateson P. (2007) *Measuring Behaviour: An Introductory Guide,* 3rd ed. Cambridge University Press.

Masi S. (2011) Differences in gorilla nettle-feeding between captivity and the wild: Local traditions, species typical behaviors or merely the result of nutritional deficiencies? *Animal Cognition* 14(6): 921.

Masi S., Cipolletta C., Robbins M.M. (2009) Western lowland gorillas (*Gorilla gorilla gorilla*) change their activity patterns in response to frugivory. *American Journal of Primatology* 71(2): 91–100.

Mason G., Mendl M.T. (1993) Why is there no simple way of measuring animal welfare? *Animal Welfare* 2: 301–319.

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.

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.

Miller L.J., Vicino G.A., Sheftel J., Lauderdale L.K. (2020) Behavioral diversity as a potential indicator of positive animal welfare. *Animals* 10(7): 1211.

Ohl F., van der Staay F.J. (2012) Animal welfare: At the interface between science and society. *The Veterinary Journal* 192(1): 13–19.

Powell D.M., Watters J.V. (2017) The evolution of the animal welfare movement in US zoos and aquariums. *Der Zoologische Garten* 86(1–6): 219–234.

Racevska E., Hill C.M. (2017) Personality and social dynamics of zoo-housed western lowland gorillas (*Gorilla gorilla gorilla*). *Journal of Zoo and Aquarium Research* 5(3): 116–122.

Ross S.R. (2006) Issues of choice and control in the behaviour of a pair of captive polar bears (*Ursus maritimus*). *Behavioural Processes* 73(1): 117–120.

Ross S.R., Calcutt S., Schapiro S.J., Hau J. (2011) Space use selectivity by chimpanzees and gorillas in an indoor-outdoor enclosure. *American Journal of Primatology* 73(2): 197–208.

Rousing T., Wemelsfelder F. (2006) Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. *Applied Animal Behaviour Science* 101(1–2): 40–53.

Sarfaty A., Margulis S.W., Atsalis S. (2012) Effects of combination birth control on estrous behavior in captive western lowland gorillas, *Gorilla gorilla gorilla. Zoo Biology* 31(3): 350–361.

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.

Stevenson-Hinde J., Stillwell-Barnes R., Zunz M. (1980) Subjective assessment of rhesus monkeys over four successive years. *Primates* 21(1): 66–82.

Stoinski T.S., Lukas K.E., Kuhar C.W., Maple T.L. (2004) Factors influencing the formation and maintenance of all-male gorilla groups in captivity. *Zoo Biology* 23(3): 189–203.

Tetley C.L., O'Hara S.J. (2012) Ratings of animal personality as a tool for improving the breeding, management and welfare of zoo mammals. *Animal Welfare* 21(4): 463.

van den Berg L.M., Bionda T., Sterck E.H.M. (2018) Protocol for a longitudinal study: Monitoring the behaviour and wellbeing of intact and castrated male Western lowland gorillas (*Gorilla gorilla gorilla*) in zoos.

Vasseur E., Gibbons J., Rushen J., de Passillé A.M. (2013) Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. *Journal of Dairy Science* 96(7): 4725–4737.

Ward S.J., Sherwen S., Clark F.E. (2018) Advances in applied zoo animal welfare science. *Journal of Applied Animal Welfare Science* 21(sup1): 23–33.

Watters J.V. (2014) Searching for behavioral indicators of welfare in zoos: Uncovering anticipatory behavior. *Zoo Biology* 33(4): 251–256.

Watts D.P. (1988) Environmental influences on mountain gorilla time budgets. *American Journal of Primatology* 15(3): 195–211.

Webb J.L., Crawley J.A., Seltmann M.W., Liehrmann O., Hemmings N., Nyein U.K., Aung H.H., Htut W., Lummaa V., Lahdenperä M. (2020) Evaluating the reliability of non-specialist observers in the behavioural assessment of semi-captive Asian elephant welfare. *Animals* 10(1): 167.

Wemelsfelder F., Hunter T.E., Mendl M.T., Lawrence A.B. (2001) Assessing the 'whole animal': A free choice profiling approach. *Animal Behaviour* 2(62): 209–220.

Whay H.R. (2007) The journey to animal welfare improvement. *Animal Welfare* 16(2): 117–122.

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

Whitham J.C., Wielebnowski N. (2013) New directions for zoo animal welfare science. A*pplied Animal Behaviour Science* 147(3–4): 247–260.

Wielebnowski N.C. (1999) Behavioral differences as predictors of breeding status in captive cheetahs. *Zoo Biology* 18(4): 335–349.

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.

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.

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.