Developing and validating attention bias tools for assessing emotion in animals: a worked example with *Macaca mulatta*.

Emmeline RI Howarth [a,b,*], Caralyn Kemp[a,c], Harriet R Thatcher[a,d], Isabelle D Szott[a], David Farningham[e], Claire L Witham[e,f], Amanda Holmes[g], Stuart Semple[g], Emily J Bethell[a]

[a] Research Centre in Brain and Behaviour, School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool, L3 3AF, UK

[b] Department of Animal Science, University Centre Myerscough, St Michael's Rd, Preston, PR3 0RY, UK

[c] School of Environmental and Animal Sciences, Unitec Institute of Technology, Auckland, 1025, New Zealand

[d] School of Biomedical Sciences, University of Edinburgh, Drummond Street, Edinburgh, EH8 9XP, UK

[e] Medical Research Council Harwell Unit, Centre for Macaques, Salisbury, SP4 0JQ, UK

[f] Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

[g] Centre for Research in Evolutionary, Social and Interdisciplinary Anthropology, University of Roehampton, London, SW15 4JD, UK

*Corresponding authors

E-mail: ehowarth@myerscough.ac.uk, telephone: +44 (0) 1995 642222 (ext. 2001)

'Declarations of interest: none'.

Author contribution statement: For Study 1, EB, AH and SS contributed to conception, and DF and CK additionally contributed to the design. CK, HT and IS collected data. For Study 2, EB, EH and CW conceived and designed the study. EH and CW built the apparatus and EH collected the data. EB

wrote the first draft of the manuscript and EB and EH wrote the final version. EB and EH conducted the statistical analysis and EH created the figures. All authors contributed to the manuscript revision and read and approved the submitted version.

This research was supported by NC3Rs grant NC/L000539/1 to EJ Bethell. ERI Howarth was supported by an LJMU PhD studentship.

Abstract

Attention bias is a new method for assessing animal emotion that has shown promising results in several animal species. Attention bias describes a tendency to preferentially attend to emotional cues compared to neutral cues, and this is influenced by underlying emotion. Because attention bias can be assessed through looking behaviour it may have broader application than other cognitive bias tasks which often require intensive periods of operant training. To date, there have been few published protocols. It is important in the early days of this new field that we develop standardised and sensitive tools and validate experimental protocols to ensure best practice. Protocols for two preferential-looking attention bias tasks are detailed: a manual task (using freely available low-cost materials: Study 1) and an automated task (requiring more expensive equipment and some programming: Study 2). Tasks were tested with 109 socially housed rhesus macaques, *Macaca mulatta,* who had been trained to sit by a target, but who had received no other training. Both tasks involved showing animals emotional face pairs (threat-neutral), filming duration of looking towards either face in the pair, and subsequent blind coding of video for duration of looking at either face. Three measures of attention to emotional faces were examined: time spent looking at the threat face, total time looking at the threat-neutral face pair overall, and attention bias difference score calculated as time spent looking at the neutral face subtracted from time spent looking at the threat face. Influence of five experimental design features on attention measures was assessed: trial number, stimulus ID, previous testing experience, time of day and visual field; as were life history variables: sex, age, and social rank. Both tasks showed sensitivity to signal (effect sizes = 0.04-0.31;

repeatabilities = 0-0.26), with reproducibility between tasks (0.15 - 0.63). All five design features had an effect on at least one measure of social attention. Trial number had a significant impact on all three measures. Agreement in raw measures was low for all three measures indicating a limit to standardisation between the two tasks. Additionally, we found evidence for stability in social attention over several years. The attention bias method shows promise for further development of standardised protocols for use with animals and we provide recommendations for future method development.

Keywords

Animal welfare, attention bias, cognitive measures, primate, rhesus macaques

1. Introduction

Attention bias (AB) methods have been proposed as a novel measure of animal emotion (Paul et al., 2005; Mendl et al., 2009; Bethell et al., 2012; Crump et al., 2018). This is based on an established human literature showing that anxiety is associated with biased attention towards threatening stimuli (Macleod et al., 1986; Bar-Haim et al., 2007). The influence of emotion on attention is accepted as an adaptive survival mechanism characteristic (LeDoux, 1996; Öhman & Mineka, 2001). AB theory and methods should therefore be generalisable across species (Paul et al., 2005). To date, AB tasks have been applied to assess emotion in a range of species (primates: Bethell et al., 2012; Marzouki et al., 2014; Allritz et al., 2016; Boggiani et al., 2018; Morin et al., 2019; birds: Brilot et al., 2009; 2012; Cussen & Mench, 2014; Campbell et al., 2019ab; sheep: Vögeli et al., 2014; Lee et al., 2016; Raoult et al., 2017; McBride & Morton, 2018; Monk et al., 2018; 2019ab; Raoult & Gygaz, 2018; 2019; cattle: Lee et al., 2018; pigs: Luo et al., 2019; rats: Parker et al., 2014). Emerging trends suggest that AB methods may provide valuable insight into animal emotion (Mendl et al., 2009; Crump et al., 2018).

To progress the field, standardised methodological and analytical approaches need to be established. This requires published protocols with worked data for the design and execution of AB tasks. Methods to assess AB in animals will vary to accommodate species' sensory adaptations, but key design aspects will likely be generalisable across species. For species reliant on visual cues, preferential looking tasks have been successfully adapted to assess emotion (e.g. primates: Bethell et al., 2012; sheep: Lee et al., 2016; cattle: Lee et al., 2018). Bethell et al. (2012) showed threat-neutral face pairs to rhesus macaques (*Macaca mulatta*) and measured duration of looking towards each face. Six threat-neutral trials were conducted in each of two conditions: baseline and 'stress'. At baseline monkeys looked more at threat faces than neutral faces (AB towards threat) but following a stressor, they became avoidant of threat faces relative to neutral faces. Lee et al. (2016) adapted the preferential looking task for use with sheep, simultaneously presenting a food bucket in the centre of an arena (positive stimulus) and a dog to one side of the arena (threat stimulus), for a single trial per sheep. Sheep who had received an anxiogenic spent significantly more time looking towards the dog than did control or anxiolytic-treated groups. Lee et al. (2018) used an almost identical single-trial protocol with cattle and found a similar effect of anxiogenic administration on attention towards the location of the dog. Thus, measures of looking behaviour have been shown to vary with emotion in several species.

A tool in development must satisfy several criteria before it is considered ready for general use (Bartlett & Frost, 2008; Kilkenny et al., 2010; see Table 1 for **Glossary**). The tool must be **sensitive** to the **signal** of interest (Bland & Altman, 1986) and provide high **repeatability** for multiple readings under identical conditions (Bartlett & Frost, 2008). Cost-effectiveness, ease of use and accessibility will increase uptake (Arthanat et al., 2009) and allow for improved intra- and inter-researcher **reliability** (Stemler, 2004). The tool must be **generalisable** across facilities, species, and between individuals, to reduce bias in use and/or animal selection (Kilkenny et al., 2010). Stimuli should be **validated** and designed to avoid familiarisation with repeated testing (Young et al., 2016). To reduce observer bias, hardware and software must be designed so that researchers are **blind** to condition at

all stages where bias may occur (Kilkenny et al., 2010). Finally, for **standardisation** of methods (e.g. between research groups), the tool should provide **reproducible** results, preferably with good **agreement** in raw data values (Buchanan-Smith, 2006; Giavarina, 2015; Prescott et al., 2017).

Detailed here are protocols for two tasks for measuring AB to social stimuli in group-housed primates. In Study 1 data were obtained using a low-cost manual apparatus with printed card stimuli. In Study 2 data were obtained using an automated apparatus presenting digital images on screens. Influence of experimental design factors was tested to identify those that must be controlled for in design and analysis. We assess the sensitivity of each method to three measures of social attention to assess which provides the most reliable signal for future studies. In addition, we tested a subset of animals using both methods to assess reproducibility of results between the two approaches (Farrell et al., 2003).

2. Methods

2.1 Animals and housing

Data were collected from 109 adult rhesus macaques (*Macaca mulatta;* 94 female) housed at the Centre for Macaques, MRC Harwell Institute, UK (mean age on first day of testing = 8.78 years, range = 2.5-18.3 years). Monkeys were UK-bred from founders of Indian origin housed in social breeding groups comprising one adult male and between three and 11 related females, plus infants and juveniles. Further details can be found in Kemp et al. (2017), Witham (2015) and at: https://www.mrc.ac.uk/research/facilities-and-resources-for-researchers/mrc-centre-for-macaques/ and www.nc3rs.org.uk/macaques. Monkeys had access to food and water *ad libitum*.

2.2 Cognitive measures

Two different sets of apparatus for conducting AB preferential-looking tasks were developed and tested. Study 1 (manual task) tested a manually operated apparatus with printed card stimuli (Figure 1a). Study 2 (automated task) tested an automated apparatus with stimuli presented as digital images on two computer screens (Figure 1b). Monkeys had previously been station trained, using

positive reinforcement, to sit next to individual 'targets' in the cage room (Kemp et al., 2017). Prior to the start of testing in each Study, monkeys underwent an initial familiarisation phase during which they were stationed, one at a time, in front of the apparatus for that Study and encouraged to look towards the apparatus by presenting food rewards centrally in front of the camera. Once animals were oriented centrally, two pictures of food items were shown to familiarise the monkey with images appearing at the two locations. Monkeys were then rewarded with the food and were free to move away at any point.

2.2.1 Stimuli

Test stimuli were pairs of pictures of seven unknown male conspecifics from the Macaque Faces Stimulus Set (Witham & Bethell, 2019; Figure 2). Each picture pair contained one frontal view of an unfamiliar male macaque face with direct gaze and mouth open, baring teeth in a tense, aggressive expression (threat face) and one frontal view of the same male with the eyes and mouth closed and face relaxed in a neutral expression (neutral face). Stimulus construction is detailed in Bethell et al. (2012).

In addition to the threat-neutral face pairs used during testing, 'filler stimuli' using presumably pleasant or neutral images for macaques were created. The inclusion of positive or neutral pictures was intended to reduce the likelihood of macaques developing a negative association with the apparatus. Filler stimuli included colour images of fruit and vegetables which the macaques are familiar with and presumably find interesting to look at (Waitt & Buchanan-Smith, 2006).

2.2.2 Study 1 (manual task)

Sixty-six female *M. mulatta* (mean age = 11.31 years, range = 2-18 years) participated in Study 1 (manual: Figure 1a). Study 1 was conducted by CK, HT and IS between 2014 and 2015 and some of the monkeys had taken part in a previous study using the same apparatus. Stimuli were printed on high quality photographic paper using a Konica High Chroma printer. The printer used was calibrated at regular intervals to produce similar levels of colour output, and the same printer was used each

time to eliminate any variation in colour output between printers. New stimuli were printed every three weeks and kept in dark conditions when not in use to avoid fading and loss of colour in images over time. Each picture measured 21cm x 28cm, thereby taking up 19 x 23 degrees of visual angle at a 60cm viewing distance.

The manual apparatus stood 2020mm tall (from floor to the top of the pole used to support the sliding framework) and 1210mm wide. The framework was adjustable to two heights that corresponded to the middle and upper levels in the cage room where monkeys would station in their preferred locations (middle level: stimuli 1150mm above ground; top level: stimuli 1780mm above ground). The apparatus was designed to look symmetrical from the monkey's perspective. A slide mechanism attached to both occluders and operated by a lever at the back revealed the two stimuli simultaneously. A Panasonic HC-V520 video camera was positioned centrally between the stimuli in front of the framework to film each monkey's direction of gaze to the stimuli.

Once a monkey was stationed by their target and looking centrally at the apparatus, the presenter moved the lever to open the occluders and reveal the stimuli, and at the same time announced 'open' as a verbal signal to indicate the start of the trial on the video. The stimuli were presented for three seconds and the participant monkey's gaze was filmed. The presenter then announced 'close' after 3 seconds as a verbal signal that the presentation of stimuli was over and the occluders were shut.

2.2.3 Study 2 (automated task)

Forty-three *M.mulatta,* participated in Study 2 (automated apparatus, n=27F, mean age = 6.65 years, range = 3-11 years; n=16M, mean age = 8.51 years, range = 3-16 years). Monkeys were naive to testing and had not taken part in Study 1. Study 2 was conducted by EH in 2018. Stimuli were digital jpeg files presented on two Eyoyo 8-inch TFT LCD Colour Video Monitors. Monitors were connected via an HDMI and a UGREEN USB to HDMI external video card to an HP ENVY 15-ah150na laptop computer. Each stimulus on the screen measured 10.2cm x 18cm, thereby taking up 9.72 x 17.06 degrees of visual angle at a 60cm viewing distance. A MATLAB program displayed the face and filler

stimuli. Each face pair was numbered in a pseudo-random manner so that the researcher conducting the trial was blind to the side of the threat face. A Sony HD video camera, mounted on a T-bar tripod and positioned equidistant between the monitors, was used to record the macaques' eye movements to the stimuli that appeared simultaneously on the adjacent monitors. At the start of the trial the researcher selected the pre-determined number for that monkey from a drop-down list in MATLAB. The face stimuli appeared simultaneously on the two screens for three seconds followed by an inter-trial interval (three seconds of black screen) and then a pair of filler stimuli were automatically presented for three seconds. A Bush SP-925 Bluetooth speaker positioned centrally at the top of the apparatus made an audible beep at stimulus onset and offset to identify the start and finish of each trial on the video. Most macaques were recorded while on the middle level of the cage room enclosure. For macaques that preferred to station on the top level, the tripod could be adjusted so that the monitors and the camera could be moved up to be in line with their eyes.

We made an additional modification to the protocol for Study 2 to reduce possible impact of the researcher. We placed a black curtain behind the apparatus and in front of the researcher so that monkeys could only see the two screens and video camera. The researcher could therefore view the monkey via the open LCD viewing screen on the camera, without looking directly at the monkey.

2.2.4 General procedure

In Study 1 (manual task), testing was conducted daily (Tuesday - Friday) for 1 week (trials 1 – 4). A subset of 18 monkeys were tested during a second week (trials 5-8) approximately 3 months later. In Study 2 testing was initially conducted daily (Tuesday - Friday) for 1 week (trials 1 – 4). A subset of 31 monkeys subsequently took part in addition trials (trials 5-13) either daily or weekly depending on accessibility. Monkeys in the same social group were always tested on the same days. All cognitive testing occurred between 9:00 and 16:30.

For all trials, we aimed to test all monkeys on days in which no identifiable major stressors had occurred. In cases where unexpected disruptions occurred and a monkey did not voluntarily come

forward to the target to take part in testing, the trial was delayed until the next available day when the monkey stationed. Some monkeys who took part in Study 1 had previously taken part in a larger study investigating the relationship between emotional state and attention to social threat, in which they had been tested during weeks when veterinary health checks took place (presumed to be stressful). We identified these monkeys as having 'prior experience' of the test, as opposed to 'naïve' monkeys who had not previously taken part in testing. All naïve monkeys had previously, but not recently, undergone the veterinary examination.

2.3 Video coding

Video was blind coded for direction and duration of eye gaze towards the left and right picture locations during each trial by CK, HT, IS and EH. Initially, 205 trials were double-blind-coded to assess inter-observer reliability. For Study 1 coding was conducted using JWatcher +Video V1.0 (Blumstein et al., 2000). Inter-observer reliability for coding was high for CK with HT (k=0.87), and HT and IZ (k=0.84). For Study 2 coding was conducted by EH using Behavioral Observation Research Interactive Software (BORIS; Friard & Gamba, 2016). Interobserver reliability was assessed between EH and CK (k=0.85). Once coded for direction and duration of gaze, data for each trial was matched with records for location of the threat face (left/right).

2.4 Data preparation

Data comprised three continuous response variables: time spent looking at the threat face per 3 second trial (THR; range = 0<3000ms), total time spent looking at the threat-neutral face pair overall per trial (TT; range = 0<3000ms) and AB difference score (ABD score). ABD score was calculated as [THR – NEUT] for each trial, providing a difference score (range = -3000ms to +3000ms) where negative values indicated more time looking towards the neutral face and positive values indicated more time looking towards the threat face.

To reduce the potential influence of outliers, we removed trials for which a disruption had occurred within a preceding time period but where the animal had still engaged with testing. Disruptions may

be associated with increased stress, which our previous work has shown influences attention bias to threat (Bethell et al., 2012). Disruptions included: receiving treatment for chronic illness, a change to group membership in the preceding week, injury in the preceding 48 hours, cleaning of housing in the last 24 hours and giving birth in the previous 24 hours.

2.5 Statistical Analysis

Statistical analyses were conducted in R v. 3.4.3 (R Core Team, 2019). A maximal model was built for each response variable for each of Study 1 and Study 2. All predictor variables were initially assessed to ensure none correlated above 0.4 (which could result in collinearity: Crawley, 2012). Response variables were visually inspected for their distribution and transformed to obtain more normal distributions when necessary (this was the case for THR and TL). Appropriate transformations were identified using Tukey's Ladder of Transformation (Tukey, 1977) to extract an appropriate lambda for transformation. Covariates were scaled using a z-transformation to a mean of zero and a standard deviation of one. Scaling continuous variables provides more comparable estimates for interpretation of model output (Aiken & West, 1991; Schielzeth, 2010). The unit of analysis was a single trial, and participant monkey identity was entered as a random factor in all models.

2.5.1 Variables

A summary of variables entered into the full model for each response variable is given in Table 2. In Study 1 (manual task) the full model for each response variable included five predictor variables associated with the experimental design, and two life-history variables. Factors were: visual field (L/R), stimulus ID (1-7), previous experience (whether monkeys had previously taken part in a separate study testing AB following a veterinary inspection, or not) and social rank (high, mid, low). Covariates were trial number (range = from 1-8 in Study 1), time of day (six one-hour time blocks from 09:00 to 15:00 in Study 1), and age (range = from 2-18 years in Study 1). In Study 2 (automated task) the full model included the same predictor variables as in Study 1 excluding: previous experience (as all monkeys in Study 2 were naïve to testing) and social rank (as this was not

equivalent between males and females). In addition, sex was included as a control variable (as males were tested in Study 2, but not Study 1) and whether cleaning had occurred in the preceding 24 hours (since this occurred often in Study 2, but not Study 1).

2.5.2 Model selection

Data were analysed using linear mixed models (LMMs) using the R package 'lme4' version 1.1-15 (Bates et al., 2015). Model stability was assessed by visually inspecting qq-plots and histograms of residuals and running influence diagnostics to identify any influential cases. Where the model was stable, predictor variables with the greatest P value were removed in a stepwise manner (Crawley, 2012) to attain a final model comprising only predictor variables with P<=0.10. We set a conservative criterion for retaining fixed effects in final models at P<=0.10 because we were interested in identifying factors which should be considered in the design and analysis of future AB studies. The 'anova' function was used to compare the fit of the final model against the null model (a model retaining the random effect, but with all fixed effects removed and an intercept of 1 specified) and the final model was accepted only if it provided a significantly better fit than the null at P<0.05 (which it did in all cases). For ABD score there were singularity warnings in both Study 1 and Study 2 for the full model indicating poor model stability. Outlier cases were identified by running a linear model (lm: the full model but with no random effect) and plotting residuals against fitted values. Outlier data points were then removed on each run until the singularity warning was resolved. Finally, we used the 'MuMin' package (Barton, 2015) to identify the conditional effect sizes for each final model (R^2). We calculated repeatability of the signal from the variance components extracted from the final model using the 'repR' package (Stoffel et al., 2017). Plots show untransformed values to aid interpretation. We present means ± 1SD unless otherwise stated. SD is presented where we were interested in variability in measurements (e.g. the effect of trial number within each study where each data point represents a single measurement). SEM is presented where we were interested in variability between individuals' mean measurements (e.g. in Study 3 where we compare means between study 1 and Study 2).

2.6 Study 3: Method comparison (Standardisation between tasks)

Reproducibility of results between the two methods was assessed for our three measures of attention using a subset of the data from Study 1 (n=18 monkeys). These 18 females were additionally tested using the automated apparatus as described in Study 2 above (but their data were not included in the analysis of Study 2 to avoid confounding effects of prior experience with a different apparatus). A typical approach for assessing reproducibility between methods is to calculate an intra-class coefficient (Koo & Li, 2016). In the present analysis it was not possible to control for confounds arising from all monkeys taking part with the manual apparatus before the automated apparatus (e.g. age, trial number and prior experience were all greater for Study 2 than Study 1). Nor was it feasible to control for stimulus ID, which varied in trial number on which it was shown between the two studies. Therefore, to control for these possible sources of within-individual variability, a mean value for each response measure was calculated per monkey for each study. We constructed a LMM for each response variable, including 'Task' (manual, automated) as the only fixed factor and 'animal ID' as a random factor. Models were assessed for stability as described above and no issues were found. Reproducibility between methods was assessed using the same calculation used for repeatability, as in this model the two are statistically equivalent. Finally, we assessed agreement by comparing the model containing the fixed factor against the null model; good agreement in values between the two methods could be assumed if the fixed factor did not improve model fit beyond the null model.

2.7 Ethical statement

Protocols were developed following discussion with the facility Home Office Inspector (Nov 2011) and carried out in accordance with ethical guidelines for work with non-human primates (NC3Rs, 2006; 2015). Approval for Study 1 was granted by the Medical Research Council Animal Welfare and Ethical Review Body (AWERB) in 2014, Roehampton University Ethics Committee (approval #LSC 14/113) and the LJMU ethics panel (approval #EB/2014-1). Approval for Study 2 was granted by the MRC AWERB in November 2017, and the LJMU ethics panel (approval #EB_EH/2017-5). Animal

health was monitored daily by the care staff, and annually with a full veterinary examination. Methods and results are reported according to the ARRIVE guidelines (Kilkenny et al., 2010).

3. Results

3.1 Study 1 (manual task)

Sixty-six females (mean age = 7.14 years, range = 2-13 years) housed in 16 social groups completed a total of 336 trials. We removed one trial due to the female recently giving birth and three trials where cleaning had occurred in the previous 24 hours (1.19% of the data removed). This resulted in 332 trials for analysis of THR and TL (mean = 5 trials per monkey, range = 4-8). For ABD score there was a singularity issue which was resolved following removal of 3 outlier cases (trial#1 from one monkey and trial#3 from two monkeys; 0.90% of the data removed). This resulted in 329 trials from 66 monkeys included in the analysis of ABD score (mean = 5 trials per monkey, range = 3-8).

Monkeys spent more time, on average, looking at the threat face (mean = 1110.36ms ± 709.26ms, range = 356.50–2211.50ms) than the neutral face (mean = 784.11ms ± 554.96ms, range = 144.50 - 1547.00ms). Consequently, ABD scores revealed overall vigilance towards threat (mean = 326.26 ± 966.95, range = -772.50–1629.25).

Model output for Study 1 is shown in Table 3. For THR and TL, trial number was the only factor retained in the final models (THR: LRT, X=21.38, df=1, P<0.001; TL: LRT, X=28.09, df=1, P<0.001; Figure 3a,b); mean time spent looking at the threat face declined from 1314ms ± 702ms on trial 1 to 643ms ± 620ms on trial 8. Mean total time spent looking at face pairs declined over trials from 2172ms ± 776ms on trial 1 to 1339ms ± 781ms on trial 8. For ABD score, trial number, previous experience and visual field were retained in the final model (trial number: LRT, X=6.05, df=1, P=0.014; previous experience: LRT, X=3.68, df=1, P=0.055; Visual field: LRT, X=2.60, df=1, P=0.11; Figure 3c,d). There was a negative relationship between ABD score, trial number and previous experience. Monkeys who had previous experience of testing after the veterinary examination had

greater ABD scores (mean = 480ms ± 986ms) than monkeys who were naïve to testing (mean = 261ms ± 953ms). Finally, there was a non-significant trend for monkeys to have lower ABD scores when the threat face was presented to the left visual field (mean = 273ms ± 1029ms) compared to when it was presented on the right (mean = 380ms ± 899ms). We conducted a post-hoc analysis to explore the main effects of previous experience and visual field by rerunning the model with these two factors entered as an interaction term. This revealed a significant interaction (LRT: X=3.88, df=1, P=0.049; Figure 3d), with the greatest difference in ABD between naïve and experience monkeys with the threat face presented on the left. The greatest ABD scores were seen in monkeys with previous experience when the threat face was presented on the left (mean ± SEM = 612ms ± 152ms), and smallest ABD scores in monkeys with no previous experience when the threat face was presented on the left (mean ± SEM = 132ms ± 90ms).

R^2 (the effect size, or coefficient of determination explaining the amount of variance in the data explained in each model) was larger when random as well as fixed effects were considered (THR: 0.17, TL: 0.31, ABD: 0.04). Repeatability was significant for THR (R=0.12, CI = 0.03 - 0.22, P= 0.005) and TL (R= 0.26, CI = 0.14 - 0.38, P <0.001) but not for ABD (R= 0.00).

3.2 Study 2 (automated task)

Forty-three monkeys (mean age = 8.03 years, range = 3–16 years; 16 males) housed in 13 social groups, and who were naive to AB testing, completed Study 2 (326 trials). Four trials were removed due to chronic illness, resulting in 322 trials from 43 monkeys for analysis of THR and TL (mean = 7.5 trials per monkey, range = 2-13; 1.23% of the data removed). For ABD score 22 outlier data points (6.83% of the data) were removed to resolve a stability issue in the model. We inspected the removed data to identify any common factors, finding nothing obvious (although males on cleaning days, one female and stimulus ID #5 may be over-represented). This resulted in 300 trials from 43 monkeys for analysis of ABD score (mean = 6.7 trials per monkey, range = 2-13).

348 Monkeys spent more time, on average per monkey, looking at the threat face (mean = 663.90ms ±
349 589.52ms, range = 20.50–1274.00ms) than the neutral face (mean = 584.03ms ± 489.19ms, range =
350 41.00-1379.14ms). Consequently, monkeys' average ABD scores revealed overall vigilance towards
351 threat (mean = 79.87 ± 791.74, range = -934.43–534.92.25).

352 Model outputs are shown in Table 3. For THR, no factors explained the variance better than the null
353 model. For TL, the factors retained in the final model were: trial number (LRT, X=4.07, df=1, P=0.044;
354 Figure 4a), time of day (LRT, X=4.12, df=1, P=0.042; Figure 4b) and sex (LRT, X=2.68, df=1, P=0.10;
355 Figure 4c). Total looking time at face pairs declined from trial 1 (mean = 1252ms ± 794ms) to trial 12
356 (mean = 842ms ± 539ms; Figure 4a) and increased between 09:00 (mean ± SEM = 963ms ± 88ms)
357 and 15:00 (mean ± SEM = 1459ms ± 215ms; Figure 4b). Males tended to have higher TL scores than
358 females (male mean ± SEM = 1417ms ± 67ms; female mean ± SEM = 1133 ms ± 50ms; Figure 4c). For
359 ABD score, stimulusID was the only factor retained in the final model (LRT, X=15.94, df=6, P=0.014;
360 Figure 4d); monkeys had greatest AB for stimulus ID #3 (mean ± SEM = 366ms ± 117ms) and lowest
361 AB scores for stimulus ID #5 (mean ± SEM = -51ms ± 74ms) and #6 (mean ± SEM = -65ms ± 90ms).

362 Effect size (R2) for each response variable was comparable to values obtained for Study 1 (THR: 0.15;
363 TL: 0.26; ABD: 0.10). Repeatability was significant for THR (R=0.12, CI = 0.04 - 0.23, P< 0.001) and TL
364 (R= 0.24, CI = 0.11 - 0.36, P <0.001) but not for ABD (R= 0.02).

365 **3.3 Study 3: Reproducibility: a method comparison of manual versus automated tasks**

366 Eighteen females (mean age = 9.08 years, range = 4.7-14.9 years) housed in nine social groups who
367 had taken part in Study 1 (manual task) were additionally tested with the automated apparatus
368 (following the protocol, but not included in the analysis, for Study 2), completing a total of 152 trials
369 (manual apparatus = 72 trials, mean = 4 trials per monkey, range = 2-6; automated apparatus = 80
370 trials, mean = 4 trials per monkey, range = 4-8). A mean value per monkey per task was calculated
371 providing a data set of 36 data points per response variable (two data points per monkey).

Linear mixed models including Task (manual or automated) as the only fixed factor and animal ID as a random factor were conducted to assess reproducibility of the two methods. These revealed significant reproducibility between the two tasks for THR (R^2=0.63, CI=0.24-0.85, P=0.001) and TL (R^2=0.39, CI=0-0.73, P=0.04), but not for ABD score (R^2=0.15, CI=0-0.58, P=0.27: Table 3). Agreement between tasks was low as the full model including Task as a fixed effect was a significantly better fit than the null model for all three response variables (THR: LRT, X^2=41.11, df=1, P<0.001; TL: LRT, X^2=39.80, P<0.001; ABD: LRT, X^2=11.07, df=1, P<0.001). All 18 monkeys had higher mean values for Study 1 than when later tested with the automated apparatus for THR and TL (THR: Study 1 mean = 1084ms, range = 0-2882ms, Study 2 mean = 537ms, range = 0-2011ms; TL: Study 1 mean = 1822ms, range = 0-3000ms, Study 2 mean = 1005ms, range = 0-2511ms; Figure 5). Sixteen of the 18 monkeys had greater ABD scores in Study 1 (group mean ± SEM = 426ms ± 89ms) than when later tested with the automated apparatus (group mean = 52ms ± 53ms), seven of whom switched from vigilance to threat in Study 1 (positive ABD scores) to avoidance of threat when later tested with the automated apparatus (negative ABD scores).

4. Discussion

Attention bias is a new method for assessing animal emotion and wellbeing with great potential due to reduced need for animal training. However, there is a paucity of protocols for designing studies with animals. Hence, it is important in this emerging field to ensure methods are robust and validated. The present study tested the efficacy of a manual and an automated apparatus to measure attention bias to emotional visual stimuli in a relatively large sample of 109 socially housed rhesus macaques. Protocols for two tasks are detailed and assessed for sensitivity to three measures of social attention: attention to threat, attention to social information and attention bias for threat (over non-threat). We also tested the influence of aspects of experimental design. Both tasks were sensitive to signals of social attention, and trial number influenced all three looking measures revealing an effect of repeated testing. Prior experience of testing, time of day and stimulus ID each

influenced one of the attention measures, with suggestive evidence for effects of visual field and sex. A method comparison for monkeys who were tested using both tasks revealed significant reproducibility for distinguishing monkeys who spent more time looking at threat faces from those who spent less time looking at threat faces. Agreement for absolute values was low between the two tasks. We discuss our findings with respect to design and analysis for future studies.

Sensitivity and repeatability: signal detection is equivalent between tasks, but not between measures of social attention

Sensitivity to signal was evidenced if one or more of the following criteria were met: obtaining good inter-observer reliability in coding direction and duration of looking from the video; evidence for an AB towards threat; and significant within-individual repeatability. For both tasks, eye-gaze towards the two stimuli was reliably blind-coded from video by two or more independent coders. This supports the use of video as an accessible and reliable, but time consuming, method for collecting looking time data as has been previously reported for a number of species (e.g. humans: Hedger et al., 2019; primates: Bethell et al., 2012; sheep: Lee et al., 2016; birds: Campbell et al., 2019b). Both tasks revealed an overall AB towards threat faces compared to neutral faces, corroborating previous studies of AB in animals finding a similar general bias towards threat (primates: Bethell et al 2012; Mandalaywala et al., 2017; Boggiani et al., 2018; sheep: Lee et al., 2016; Raoult & Gygaz, 2019; birds: Brilot et al., 2012; cattle: Lee et al., 2018; pigs: Luo et al., 2019; review: Crump et al., 2018), and as is the general pattern for humans (Bar-Haim et al., 2007). Finally, there was significant within-individual repeatability for looking towards threat faces and for total duration of looking towards both faces, but not for ABD score. This indicates detectable individual differences (traits) in duration of looking towards faces. For all three measures of social attention, repeatabilities were within the range reported in the animal behaviour literature (e.g. Bell et al., 2009, reported on a range of behaviours across species, mean R = 0.37, range = -1-1). Greater repeatability for duration of looking towards threat faces and at threat-neutral face pairs, compared with near-zero repeatability for ABD

scores, suggests either that ABD scores are more sensitive to transient emotional states and may be useful for detecting shifts in emotion (as suggested by Bethell et al., 2012), or that the calculation introduces additional noise, creating a less reliable measure. This requires further study including longitudinal studies and further examination of the individual characteristics (genetic and developmental) and environmental factors that impact social attention.

While both the manual and automated tasks were sensitive to signal, there were some notable differences. Looking times towards stimuli were longer and ABD scores were greater with the manual apparatus than the automated apparatus. It is possible that the movement of the occluders and verbal cue of the researcher at the start of each trial in Study 1 may have increased the salience of the apparatus and stimuli, and possibly had a startling effect on some monkeys, influencing spatial orienting (Balaban, 1995; Lang, 1995; Hess et al., 2007; Irwin, 2011; Lane et al., 2013). Shifts in emotion and arousal are known to influence AB (primates: Bethell et al., 2012; humans: Bar-Haim et al., 2007). Automated methods are likely to reduce distracting movement and therefore noise in data, increasing sensitivity to signal (Tipper et al., 1998; Mandillo et al., 2008; Bains et al., 2018).

Reproducibility between tasks reveals stable individual differences over time, with low agreement

Analysis of data from monkeys that were tested using both methods revealed reproducibility of findings with respect to ranking monkeys for social attention, with low agreement in raw values. Monkeys with greatest looking time towards threat faces and threat-neutral face pairs when tested with the manual apparatus in 2013-2014, also had the greatest looking time towards faces with the automated apparatus in 2018. This indicates that the two tasks provide reproducible results, and that individual differences in social attention are stable across several years in adult macaques. There was no reproducibility for ABD scores. This may be because the measure is noisy, or alternatively because it has high sensitivity to transient emotional states so that averaging data dilutes the signal and testing at disparate time points likely introduces latent confounds related to emotion state. Despite the reproducibility for two of the measures, there was low agreement in raw

values for all three measures, possibly due to the issues with sensitivity already discussed. It is likely data need to be normalised before comparison can be attempted between studies.

Experimental design and controlling for confounds

Nine variables were assessed for the possible influence on our three measures of social attention (five related to experimental design, one husbandry procedure and three life history variables). All five variables related to experimental design had a significant impact on at least one measure of social attention. There was no evidence that recent cleaning impacted results or that the three life history variables had a significant impact, although males may look at face pairs for longer than do females. We discuss each aspect of experimental design in turn.

Trial number: repeated testing is associated with reduction in all three measures of social attention, but inter-trial interval is important

Trial number had a significant effect on all three measures indicating a large influence of repeated testing. The effect of repeated testing on social attention and task performance is well established across paradigms and species (humans: Nanhoe-Mahabier et al., 2012; Denny et al., 2014; primates: Bethell et al 2019; sheep: Doyle et al 2010; rats: Spruijt 1992). Studies with humans have shown the dissipation of attentional bias effects across trials within a single testing session (e.g. Cohen et al., 1998), although more recent work indicates that emotion-specific attentional biases do not habituate rapidly (e.g., Lonsdorf et al., 2014). Studies with primates have found possible habituation effects of repeated testing (King et al., 2012), with effects being greatest over the first few trials in some cases (Bethell et al., 2019). To mitigate these effects, we conducted one trial per day on consecutive days for the first four trials in each of Study 1 and Study 2, however, effects of repeated testing were still evident suggesting an inter-trial interval of 24 hours is too short to eradicate confounds such as carry-over effects or loss of interest. In Study 2, trials 9-13 were conducted at weekly intervals and testing at this interval appears to have revived interest in stimuli, as the effect of trial number was lost for looking time towards threat and ABD score, and was positive for total

duration of looking at faces pairs. Therefore, interest in stimulus pairs was maintained when trials were conducted 7 days apart supporting weekly rather than daily testing.

Previous experience of testing (when stressed) enhances AB towards threat in the left visual field

Monkeys who had previous experience of testing when stressed as part of a larger study, showed greater AB towards threat faces in Study 1 than did monkeys who were naïve to testing. This effect was evident when the threat face was presented to the left visual field, but not the right visual field. This finding is in line with the left visual-field bias for recognition of emotional faces arising from a right hemisphere superiority for emotional face processing (e.g. humans: Borod et al., 1998; Mandal & Ambady, 2004; Najt et al., 2013; primates: Lindell, 2013), especially of negative emotional information (Ahern & Schwartz, 1985; Adolphs et al., 2001; Jansari et al., 2011). Prior experience and hemispheric effects should therefore be controlled for in study design and analysis. In the present study, it is possible that first witnessing the stimuli when stressed resulted in context dependent associative learning effects and enhanced negative value of the threat faces during the current study (Mendl, 1999; Bliss-Moreau et al., 2008; Richards et al., 2013). In some human studies images are presented on a vertical plane with a top and bottom image in addition to left and right (e.g. Eimer & Holmes, 2007). This may mitigate some lateralisation effects. Use of single stimuli may also prove a fruitful alternative (Eimer & Holmes, 2007; Winters et al., 2015).

Stimulus identity and sex of viewer impact attention bias

We used seven threat-neutral face pairs compiled from colour photographs of males from the Macaque Faces stimulus set (Witham & Bethell, 2019). When presented in digital format we found a significant influence of stimulus ID on ABD score. Due to using photographs of real animals it is likely the stimulus pairs varied from each other in luminance, contrast energy, colour and brightness, as well as emotional intensity, dominance, attractiveness, age, orientation and even degree of head tilt, all of which can influence attention to faces (Hess et al., 2007; Palumbo et al., 2017; Waitt & Buchanan-Smith 2006). Human studies use large numbers of images from picture libraries to reduce

497 influence of individual stimuli. These studies have revealed variation in AB for angry, disgust and
498 pain facial expressions (Schofield et al., 2013; Hommer et al., 2014; Heathcote et al., 2015). Large
499 picture databases are not commonly available for animal studies, although picture libraries (e.g.
500 macaques: Witham & Bethell 2019) and avatars (e.g. macaques: Wilson et al., 2019) are coming
501 online. Primate work developing attentional measures of affect has largely followed the human
502 literature using pictures of emotional and neutral faces (Bethell et al., 2012; 2016; 2019; Cronin et
503 al., 2018; Crump et al., 2018). Other stimuli tested with primates include veterinary and husbandry
504 stimuli (Bethell et al., 2019; Allritz et al., 2016; Boggiani et al., 2018). Work with sheep and cattle has
505 used a dog (Lee et al., 2016; 2018) and birds have been tested using eye spots (Brilot et al., 2009).
506 Development of large stimulus sets for studies is necessary to ameliorate the potentially
507 confounding effects of individual effects.

508 Human studies have shown an interaction between sex of viewer and stimulus ID (Hess & Thiabault,
509 2009; Hess et al., 2009; Brody et al., 2012), which may further be influenced by individual
510 characteristics such as dominance motivation as well as emotion expression (Hareli et al., 2009;
511 2015). For example, women exhibit an own-gender bias in attention to faces, which is not present in
512 men (Lovén et al., 2011; Herlitz & Lovén, 2013). This gender bias in human is mirrored in primates
513 with female capuchin monkeys showing an AB towards images of female conspecifics over male
514 conspecifics while male capuchin monkeys showed no preference (Schino et al., 2020). While we
515 found no indication that attention to the stimulus IDs varied between the sexes, males tended to
516 spend more time looking at digitally presented male face pairs overall than did females. As there was
517 no effect of sex on looking time towards threat faces alone, males may have found neutral faces
518 with eyes closed more appealing to look at than did females. This is consistent with human literature
519 for sex differences in sensitivity and attention allocation to threat (Campbell & Muncer, 2017). We
520 are investigating the influence of sociosexual factors on social attention as part of a larger study.

521 **Social attention follows a circadian rhythm**

Social attention for digitally presented face pairs increased from the morning into the afternoon. Attention in humans shows a circadian rhythm (Valdez et al., 2005). It is therefore likely the effect of time of day reflects circadian changes in alertness and arousal that have been documented across animal species (e.g. cows: Niu et al., 2014; birds: Ramli & Norazlimi, 2016; primates: Kappeler & Erkert, 2003; Plant, 1981; Novak et al, 2013). Daily husbandry schedules, such as feeding and cleaning, may also influence engagement with tasks at certain times. Time should be controlled for where testing time varies.

5. Conclusion

The preferential looking attention bias method presented here meets a number of criteria for consideration as a valid tool for assessing social attention. We found sensitivity to signal varied between our measures of social attention, indicating the different measures may better fit particular research questions (for example attention bias difference score may better detect transient shifts in emotion, while attention to threat reflects more stable trait emotional characteristics). Reproducibility between tasks indicates the method should be adaptable between facilities, especially where standardised protocols and stimulus sets are shared. A testing schedule with a minimum weekly interval between trials would eliminate carry-over effects that were evident in our study from daily testing, and previous experience, time of day, stimulus ID, animal sex and hemispheric factors should all be controlled for in design in analysis. We hope the protocols and analyses presented in this paper will be useful to researchers designing and implementing future studies of social attention and attention bias, including the application of these methods for studying animal emotion.

Acknowledgements

We thank the staff at CFM, particularly Faye Peters and Sebastian Merritt who assisted with data collection.

References

Adolphs R, Jansari A & Tranel D, 2001. Hemispheric perception of emotional valence from facial expressions. *Neuropsychology*, 15(4), 516–524.

Ahern GL & Schwartz GE, 1985. Differential lateralization for positive and negative emotions in the human brain: EEG spectral analysis. *Neuropsychologia*, 23, 745–755.

Aiken LS & West SG, 1991. *Multiple regression: testing and interpreting interactions*. Sage, Newbury Park.

Allritz M, Call J & Borkenau P, 2016. How chimpanzees (*Pan troglodytes*) perform in a modified emotional Stroop task. *Animal Cognition*, 19(3), 435-449.

Arthanat S, Nochajski SM, Lenker JA, Bauer SM & Wu YWB, 2009. Measuring usability of assistive technology from a multicontextual perspective: The case of power wheelchairs. *American Journal of Occupational Therapy,* 63, 751-764.

Bains RS, Wells S, Sillito RR, Armstrong JD, Cater HL, Banks G & Nolan PM, 2018. Assessing mouse behaviour throughout the light/dark cycle using automated in-cage analysis tools. *Journal of Neuroscience Methods*, 300, 37–47.

Balaban MT, 1995. Affective influences on startle in five-month-old infants: Reactions to facial expressions of emotion. *Child Development*, 66, 28–36.

Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ & van Ijzendoorn MH, 2007. Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin,* 133, 1-24.

Bartlett JW & Frost C, 2008. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics & Gynecology*, 31, 466-475.

Barton K, 2015. Package 'MuMIn'. Version 1, 18.

Bates D, Machler M, Bolker BM & Walker SC, 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1-48.

Bell AM, Hankinson SJ & Laskowski KL, 2009. The repeatability of behaviour: a meta-analysis. *Animal Behaviour*, 77(4), 771-783.

Bethell EJ, Cassidy LC, Brockhausen RR & Pfefferle D, 2019. Toward a standardized test of fearful temperament in primates: A sensitive alternative to the Human Intruder Task for laboratory-housed Rhesus Macaques (*Macaca mulatta*). *Frontiers in Psychology*, 10, 1-17.

Bethell EJ, Holmes A, MacLarnon A & Semple S, 2012. Evidence that emotion mediates social attention in rhesus macaques. *Plos One*, 7, 1-9.

Bethell EJ, Holmes A, MacLarnon A & Semple S, 2016. Emotion evaluation and response slowing in a non-human primate: new directions for cognitive bias measures of animal emotion? *Behavioral Sciences,* (2076-328X) 6, 1-16.

Bland MJ & Altman DG, 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327, 307-310.

Bliss-Moreau E, Barrett LF, & Wright CI, 2008. Individual differences in learning the affective value of others under minimal conditions. *Emotion*, *8*(4), 479.

589 Blumstein DT, Daniel JC & Evans CS, 2000. JWatcher +Video 1.0, In: http://www.jwatcher.ucla.edu/
590 (Ed.).

591 Boggiani L, Addesi, E & Schino G, 2018. Receiving aggression triggers attention bias in tufted
592 capuchin monkeys. *Animal Behaviour,* 146, 173-180.

593 Borod JC, Cicero BA, Obler LK, Welkowitz J, Erhan HM, Santschi C, Grunwalk IS, Agosti RM & Whalen
594 JR, 1998. Right hemisphere emotional perception: Evidence across multiple channels.
595 *Neuropsychology*, 12(3), 446–458.

596 Brilot BO & Bateson M, 2012. Water bathing alters threat perception in starlings. *Biology Letters*, 8,
597 379–381.

598 Brilot BO, Normandale CL, Parkin A & Bateson M, 2009. Can we use starlings' aversion to eyespots as
599 the basis for a novel 'cognitive bias' task? *Applied Animal Behaviour Science*, 118, 182-190.

600 Brody S, Simard A & Hess U, 2012. Men's sexual activity and perceptions of the facial attractiveness
601 of unknown women. *Sexual and Relationship Therapy*, 27(4), 372-376.

602 Buchanan-Smith HM, 2006. Primates in laboratories: Standardisation, harmonisation, variation and
603 science. *ALTEX: Alternatives to Animal Experimentation*, 115-119.

604 Campbell DLM, Dickson EJ & Lee C, 2019a. Application of open field, tonic immobility, and attention
605 bias tests to hens with different ranging patterns. *PeerJ*, 7, e8122.

606 Campbell DLM, Taylor PS, Hernandez CE, Stewart M & Belson S & Lee C, 2019b. An attention bias
607 test to assess anxiety states in laying hens. *PeerJ*, 7, e7303.

608 Campbell A & Muncer S, 2017. Sex difference in awareness of threat: A meta-analysis of sex
609 differences in attentional orienting in the dot probe task. *Personality and Individual Differences*, 119,
610 181-184.

611 Cohen DJ, Eckhardt CI & Schagat KD, 1998. Attention allocation and habituation to anger-related
612 stimuli during a visual search task. *Aggressive Behavior: Official Journal of the International Society*
613 *for Research on Aggression*, 24(6), 399-409.

614 Crawley MJ, 2012. *The R Book*. Wiley.

615 Cronin KA, Bethell EJ, Jacobson SL & Ross SR, 2018. Evaluating mood changes in response to
616 anthropogenic noise with a Response-Slowing Task in three species of zoo-housed primates. *Animal*
617 *Cognition*, 5, 209-–221.

618 Crump A, Arnott G & Bethell EJ, 2018. Affect-driven attention biases as animal welfare indicators:
619 review and methods. *Animals*, 8.

620 Cussen VA & Mench JA, 2014. Personality predicts cognitive bias in captive psittacines, *Amazona*
621 *Amazonica*. *Animal Behaviour*, 89, 123–130.

622 Denny BT, Jin F, Xun L, Guerreri S, Mayson SJ, Rimsky L, New AS, Siever LJ, Koenigsberg HW, 2014.
623 Insula-amygdala functional connectivity is correlated with habituation to repeated negative images,
624 *Social Cognitive and Affective Neuroscience*, 9,1660 –1667.

625 Doyle RE, Vidal S, Hinch GN, Fisher AJ, Boissy A & Lee C, 2010. The effect of repeated testing on
626 judgement biases in sheep. *Behavioural Processes*, 83(3), 349-452.

627 Eimer M & Holmes A, 2007. Event-related brain potential correlates of emotional face processing.
628 *Neuropsychologia*, 45(1), 15-31.

629 Farrell T, Cairns M & Leslie J, 2003. Reliability and validity of two methods of three-dimensional
630 cervical volume measurement. *Ultrasound in Obstetrics & Gynecology,* 22, 49-52.

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

633 Giavarina D, 2015. Understanding Bland Altman analysis. *Biochemia Medica (Zagreb),* 25, 141-151.

634 Hareli S, David S & Hess U, 2015. The role of emotion transition for the perception of social
635 dominance and affiliation. *Cognition and Emotion*, 30(7), 1260-1270.

636 Hareli S, Shomray N & Hess U, 2009. Emotional versus neutral expressions and perceptions of social
637 dominance and submissiveness. *Emotion*, 9(3), 378-84.

638 Heathcote LC, Vervoort T, Eccleston C, Fox E, Jacobs K, Van Ryckeghem DML & Lau JYF, 2015. The
639 relationship between adolescents' pain catastrophizing and attention bias to pain faces is moderated
640 by attention control. *Pain*. 156, 1334-1341.

641 Hedger N, Garner M & Adams W, 2019. Do emotional faces capture attention, and does this depend
642 on awareness? Evidence from the visual probe paradigm. *Journal of Experimental Psychology:*
643 *Human Perception and Performance,* 45(6), 790-802.

644 Herlitz A & Lovén J, 2013. Sex differences and the own-gender bias in face recognition: A meta-
645 analytic review. *Visual Cognition*, 21, 9-10, 1306-1336.

646 Hess U, Adams RB, Grammer K & Kleck R, 2009. Face gender and emotion expression: Are angry
647 women more like men? *Journal of Vision*, 9(12), 1-8.

648 Hess U & Thibault P, 2009. Why the same expression may not mean the same when shown on
649 different faces or seen by different people. In: Tao J & Tan T (eds.) *Affective Information Processing*,
650 145-158. Springer: London.

651 Hess U, Adams RB & Kleck RE, 2007. Looking at you or looking elsewhere: The influence of head
652 orientation on the signal value of emotional facial expressions. *Motivation and Emotion*, 31(2), 137-
653 144.

654 Home Office, 2014. Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for
655 Scientific Purposes, In: Office, H. (Ed.), p. 227.

656 Hommer RE, Meyer A, Stoddard J, Connolly ME, Mogg K, Bradley BP, Pine DS, Leibenluft E & Brotman
657 MA, 2014. Attention bias to threat faces in severe mood dysregulation. *Depression & Anxiety,* (1091-
658 4269), 31, 559-565.

659 Irwin DE, 2011. Where does attention go when you blink? *Attention, Perception, & Psychophysics*,
660 73, 1374–1384.

661 Jansari A, Rodway P & Goncalve S, 2011. Identifying facial emotions: Valence specific effects and an
662 exploration of the effects of viewer gender. *Brain and Cognition*, 76, 415–423.

663 Kappeler PM & Erkert HG, 2003. On the move around the clock: correlates and determinants of
664 cathemeral activity in wild red fronted lemurs (*Eulemur fulvus rufus*). *Behavioral Ecology and*
665 *Sociobiology*, 54, 359-369.

666 Kemp C, Thatcher H, Farningham D, Witham C, MacLarnon A, Holmes A, Semple S & Bethell EJ, 2017.
667 A protocol for training group-housed rhesus macaques (Macaca mulatta) to cooperate with
668 husbandry and research procedures using positive reinforcement. *Applied Animal Behaviour Science*,
669 197, 90-100.

670 Kilkenny C, Browne WJ, Cuthill IC, Emerson M & Altman DG, 2010. Improving bioscience research
671 reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology,* 8, e1000412.

672 King HM, Kurdziel LB, Meyer JS & Lacreuse A, 2012. Effects of testosterone on attention and memory
673 for emotional stimuli in male rhesus monkeys. *Psychoneuroendocrinology*, 37(3), 396–409.

674 Koo TK & Li MY, 2016. A guideline of selecting and reporting intraclass correlation coefficients for
675 reliability research. *Journal of Chiropractic Medicine*, 15, 155-163.

Lane ST, Franklin JC & Curran PJ, 2013. Clarifying the nature of startle habituation using latent curve modeling. *International Journal of Psychophysiology*, 88, 55-63.

Lang PJB, Bradley MM & Cuthbert BN, 1997. Motivated attention: affect, activation, and action. In: PJ Lang, RF Simons, MT Balaban (Ed.), *Attention and Orienting: Sensory and Motivational Processes*, Erlbaum: Hillsdale, NJ, pp. 97-135.

LeDoux J, 1996. *The Emotional Brain*. Simon & Schister, New York

Lee C, Cafe LM, Robinson SL, Doyle RE, Lea JM, Small AH & Colditz IG, 2018. Anxiety influences attention bias but not flight speed and crush score in beef cattle. *Applied Animal Behaviour Science,* 205, 210-215.

Lee C, Verbeek E, Doyle R & Bateson M, 2016. Attention bias to threat indicates anxiety differences in sheep. *Biology Letters,* 12, 20150977.

Lindell AK, 2013. Continuities in emotion lateralization in human and non-human primates. *Frontiers in Human Neuroscience*, 7, 464.

Lonsdorf TB, Juth P, Rohde C, Schalling M & Öhman A, 2014. Attention biases and habituation of attention biases are associated with 5-HTTLPR and COMTval158met. *Cognitive, Affective, & Behavioral Neuroscience*, 14(1), 354-363.

Lovén J, Herlitz A & Rehnman J, 2011. Women's own-gender bias in face recognition memory: The role of attention at encoding. *Experimental Psychology*, 58, 333-340.

Luo L, Reimert I, de Haas EN, Kemp B & Bolhuis JE, 2019. Effects of early and later life environmental enrichment and personality on attention bias in pigs (*Sus scrofa domesticus*). *Animal Cognition*, 22, 959-972.

Macleod C, Mathews, A & Tata P, 1986. Attentional bias in emotional disorders. *Journal of Abnormal Psychology*, 95, 15-20.

Mandalaywala TM, Pertrullo LA, Parker KJ, Maestripieri D & Higham JP, 2017. Vigilance for threat accounts for inter-individual variation in physiological responses to adversity in rhesus macaques: A cognition x environment approach. *Developmental Psychobiology*, 59(8), 1031-1038.

Mandal MK & Ambady N, 2004. Laterality of facial expressions of emotion: Universal and culture-specific influences. *Behavioural Neurology*, 15, 23–34.

Mandillo S, Tucci V, Hölter SM, Meziane H, Banchaabouchi MA, Kallnik M, Lad HV, Nolan PM, Ouagazzal AM, Coghill EL, Gale K, Golini E, Jacquot S, Krezel W, Parker A, Riet F, Schneider I, Marazziti D, Auwerx J, Brown SD, Chambon P, Rosenthal N, Tocchini-Valentini G & Wurst W, 2008. Reliability robustness, and reproducibility in mouse behavioral phenotyping: a cross-laboratory study. *Physiological Genomics*, 34, 243–255.

Marzouki Y, Gullstrand J, Goujon A & Fagot J, 2014. Baboons' response speed is biased by their moods. *PLoS ONE*, 9, e102562.

McBride SD & Morton AJ, 2018. Visual attention and cognitive performance in sheep. *Applied Animal Behaviour Science,* 206, 52-58.

Mendl M, 1999. Performing under pressure: stress and cognitive function. *Applied Animal Behaviour Science*, 65(3), 221-244.

Mendl M, Burman, OHP, Parker RMA & Paul ES, 2009. Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. *Applied Animal Behaviour Science,* 118, 161-181.

Monk JE, Belson S & Lee C, 2019a. Pharmacologically-induced stress has minimal impact on judgement and attention biases in sheep. *Scientific Reports*, 9, 11446.

Monk JE, LC, Belson S, Colditz IG & Campbell DLM, 2019b. The influence of pharmacologically-induced affective states on attention bias in sheep. *PeerJ*, 7, e7033.

Monk JE, Doyle RE, Colditz IG, Belson S, Cronin GM & Lee C, 2018. Towards a more practical attention bias test to assess affective state in sheep. *Plos One*, 13, 15.

Morin EL, Howell BR, Meyer JS & Sanchez MM, 2019. Effects of early maternal care on adolescent attention bias to threat in nonhuman primates. *Developmental Cognitive Neuroscience*, 38, 100643.

Nanhoe-Mahabier W, Allum JHJ, Overeem S, Borm GF, Nijhuis LBO & Bloem BR, 2012. First trial reactions and habituation rates over successive balance perturbations in Pakinson's disease. *Neuroscience* 217, 123–129.

Najt P, Bayer U & Hausmann M, 2013. Models of hemispheric specialization in facial emotion perception—A revaluation. *Emotion*, 13(1), 159–167.

NC3Rs, 2006. Primate Accommodation, Care and Use. NC3Rs, London.

NC3Rs, 2015. The Macaque website, National Centre for the Replacement, Refinement and Reduction of animals in research, http://www.nc3rs.org.uk/macaques/.

Niu M, Ying Y, Bartell PA & Harvatine KJ, 2014. The effects of feeding time on milk production, total-tract digestibility, and daily rhythms of feeding behavior and plasma metabolites and hormones in dairy cows. *Journal of Dairy Science*, 97(12), 7764-7777.

Novak MA, Hamel AF, Kelly BJ, Dettmer AM & Meyer JS, 2013. Stress, the HPA axis, and nonhuman primate well-being: a review. *Applied Animal Behaviour Science*, 143(2-4), 135-149.

Öhman A & Mineka S, 2001. Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, 108, 483-522.

Palumbo R, Adams RB, Hess U, Kleck R & Zebrowitz L, 2017. Age and gender differences in facial attractiveness, but not emotion resemblance, contribute to age and gender stereotypes. *Frontiers in Psychology*, 8, 1704.

Parker RM, Paul ES, Burman OH, Browne WJ & Mendl M, 2014. Housing conditions affect rat responses to two types of ambiguity in a reward–Reward discrimination cognitive bias task. Behavioural Brain Research, 274, 73–83.

Paul ES, Harding EJ & Mendl M, 2005. Measuring emotional processes in animals: the utility of a cognitive approach. *Neuroscience and Biobehavioral Reviews,* 29, 469-491.

Peris-Vicente J, Esteve-Romero J & Carda-Broch S, 2015. Chapter 13 - Validation of analytical methods based on chromatographic techniques: An overview. In: Anderson JL, Berthod A, Estévez VP, Stalcup AM (Eds.), *Analytical Separation Science* (1st Edn.). Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany. 1757.

Plant TM, 1981. Time courses of concentrations of circulating gonadotropin, prolactin, testosterone, and cortisol in adult male rhesus monkeys (*Macaca mulatta*) throughout the 24 h light-dark cycle. *Biology of Reproduction*, 25, 244–252.

Prescott MJ, Langermans JA & Ragan I, 2017. Applying the 3Rs to non-human primate research: barriers and solutions. *Drug Discovery Today: Disease Models,* 23, 51-56.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramli R & Norazlimi NA, 2016. Effects of tidal states and time of day on the abundance and behavior of shorebirds utilizing tropical intertidal environment. *Journal of Animal & Plant Sciences*, 26(4), 1164-1172.

763 Raoult CMC & Gygax L, 2019. Mood induction alters attention toward negative-positive stimulus
764 pairs in sheep. *Scientific reports*, 9, 7759.

765 Raoult CMC & Gygax L, 2018. Valence and intensity of video stimuli of dogs and conspecifics in
766 sheep: Approach-avoidance, operant response, and attention. *Animals*, 8, 121.

767 Raoult CMC, Moser J & Gygax L, 2017. Mood as cumulative expectation mismatch: A test of theory
768 based on data from non-verbal cognitive bias tests. *Frontiers in Psychology*, 8.

769 Richards A, Holmes A, Pell PJ & Bethell EJ, 2013. Adapting effects of emotional expression in anxiety:
770 Evidence for an enhanced Late Positive Potential. *Social Neuroscience*, 8, 650-664.

771 Schielzeth H, 2010. Simple means to improve the interpretability of regression coefficients. *Methods*
772 *in Ecology and Evolution*, 1, 103-113.

773 Schino G, Carducci P & Truppa V, 2020. Attention to social stimuli is modulated by sex and exposure
774 time in tufted capuchin monkeys. *Animal Behaviour*, 161, 39-47.

775 Schofield CA, Inhoff AW & Coles ME, 2013. Time-course of attention biases in social phobia. *Journal*
776 *of Anxiety Disorders*, 27, 661-669.

777 Spruijt BM, 1992. Progressive decline in social attention in aging rats: An information-statistical
778 method. *Neurobiology of Aging*, 13(1), 145-151.

779 Stemler SE, 2004. A comparison of consensus, consistency, and measurement approaches to
780 estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9.

781 Stoffel MA, Nakagawa S & Schielzeth H, 2017. rptR: repeatability estimation and variance
782 decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 8(11),
783 1639–1644.

784 Tate J & Panteghinim M, 2007. Standardisation - The theory and the practice. *The Clinical Biochemist*
785 *Reviews*, 28, 93-96.

786 Tipper SP, Howard LA & Houghton G, 1998. Action-based mechanisms of attention. *Philosophical*
787 *Transactions: Biological Sciences, the Royal Society of London*, 353(1373), 1385-1393.

788 Tomonaga M & Imura T, 2015. Efficient search for a face by chimpanzees (Pan troglodytes). *Scientific*
789 *Reports,* 5, 11437.

790 Tukey J, 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

791 Valdez P, Ramírez C, García A, Talamantes J, Armijo P & Borrani J, 2005. Circadian rhythms in
792 components of attention. B*iological Rhythm Research*, 36(1–2), 57–65.

793 Vögeli S, Lutz J, Wolf M, Wechsler B & Gygax L, 2015. Valence of physical stimuli, not housing
794 conditions, affects behaviour and frontal cortical brain activity in sheep. *Behavioural Brain Research*,
795 267, 144–155.

796 Waitt C & Buchanan-Smith HM, 2006. Perceptual considerations in the use of colored photographic
797 and video stimuli to study nonhuman primate behavior. *American Journal of Primatology*, 68, 1054-
798 1067.

799 Wilson VAD, Kade C, Moeller S, Treue S, Kagan I, Fischer J, 2019. Development of a monkey avatar to
800 study social perception in macaques. *BioRxiv.* DOI:10.1101/758458.

801 Winters S, Dubuc C & Higham JP, 2015. Perspectives: the looking time experimental paradigm in
802 studies of animal visual perception and cognition. *Ethology*, 121(7), 625-640.

803 Witham CL, 2015. Automated face recognition of rhesus macaques. *Journal of Neuroscience*
804 *Methods,* 300, 157-165.

805 Witham C & Bethell EJ, 2019. Macaque Faces. *Figshare*, Dataset.
806 https://doi.org/10.6084/m9.figshare.9862586

807 Young SG, Jones IF & Claypool HM, 2016. Stimulus threat and exposure context modulate the effect
808 of mere exposure on approach behaviors. *Frontiers in Psychology*, 7.

809

810 **Tables**

811

Table 1. Glossary of terms

Agreement is the extent to which two or more methods give numerically identical values for all individuals (Bland & Altman, 1986). Agreement is different from correlation since correlation between two methods can be high while agreement in absolute values can be low.
Attention bias refers to the difference in allocation of attention towards one emotional stimulus compared to another when both are presented simultaneously. It can be measured as direction of eye-gaze (as we do here), head orientation (e.g. Bogianni et al., 2018), or more indirectly through other measures such as vigilant scanning (e.g. Allritz et al., 2016), or changes in reaction time on cognitive tasks such as the dot-probe (e.g. King et al., 2012; Tomonaga and Imura, 2015).
Blind coding occurs when a coder is unaware of the experimental conditions, in order to remove unconscious bias effects during coding. Here, video of attention bias trials was coded without knowledge of the location of the threat face (left or right).
Generalisability is whether, and how, the findings of a study are likely to translate to other species or facilities (Kilkenny et al., 2010).
Inter-observer reliability is the agreement between two or more observers in coding a measure. Here, an agreement matrix was compiled for each pair of scorers and used to record agreement in coding direction of eye-gaze in each frame of video. Reliability for each pair of coders was then calculated using Cohen's kappa which accounts for variability in reliability of coding across different categories (Cohen, 1960).
Repeatability of measurements refers to the variation in repeat measurements made on the same subject using the same method. There is an assumption that measurements are made under identical conditions by the same researcher over a short period of time (Bartlett & Frost, 2008). Here we test repeatability of measurements for individual animals when tested using each variant of the method, separately.
Reproducibility of results refers to the variation in measurements made on the same subject using the different variants of the method. There is an assumption that measurements are made within a period of time within which no change in the variable should have occurred due to other factors (Bartlett & Frost, 2008). Here we test reproducibility of tasks for 16 animals, tested at a 3-year interval.
Sensitivity refers to the ability of a protocol to detect the signal of interest, here longer looking times (bias in attention) towards one stimulus over another.

Standardisation is the practice of adopting the same methods or practices in order for results to be comparable (Tate and Panteghini, 2007). Here, we compare protocols (manual vs. automatic) in order to determine if the use of these different tasks at different facilities results in comparable results.

Validation ensures methods provide reliable, consistent, accurate and high-quality data (Peris-Vicente et al., 2015). Seven threat-neutral conspecific face pair stimuli were validated for strength of signal and their applicability for use in attention bias studies.

812

Table 2. List of variables used in Study 1 (manual) and Study 2 (automated).

Category	Variable type	Variable name	Levels	Study	Description
Experimental design	Factor	Visual field	2	1 & 2	The threat face was shown either at the left or the right location
		Stimulus ID	7	1 & 2	There were seven pairs of stimuli, each pair of a different monkey identity
		Previous experience	2	1	In Study 1, 23/66 monkeys had recent previous experience of the attention bias task following a veterinary inspection. In Study 2 all 43 monkeys were naïve to attention bias testing.
	Covariate	Trial number	na	1 & 2	Monkeys took part in one trial per day, usually on four consecutive days in a week
		Time of day	na	1 & 2	Time of day at which testing occurred, recorded in 1-hour time blocks between 09:00 and 16:00
Husbandry	Factor	Cleaning in last 24hrs	2	2	In Study 1 we removed trials where the home cage had been cleaned in the preceding 24 hours (n=3 trials). In Study 2 we included this as a control variable as it happened more often (n= 34 trials)
Life history	Factor	Rank	3	1	Monkeys were classified as high, medium or low in social rank.
		Sex	2	2	Study 1 included females only, in Study 2 we tested females and males
	Covariate	Age	na	1 & 2	Age of monkey (yrs) on day of testing

813

814

815

Table 3. Model output for Study 1 (manual; n=66) and Study 2 (automated; n=43). For reproducibility between Study 1 and Study 2 n=18.

Measure	Study	Final Model predictors	Estimate	SE	t	95%CIL	95%CIU	X2	df	P	R2m	R2c	Repeat-ability within indiv.	Repro-ducibility
THR	1	(Intercept)	44.41	1.15							0.06	0.17	0.121	0.63
		Trial number	-4.40	0.93	-4.71	-6.23	-2.55	21.38	1	**<0.001**				
	2	(Intercept)	19.27	0.73							0.02	0.15	0.12	
		-	-	-	-	-	-	-	-	-				
TL	1	(Intercept)	408.10	11.72							0.07	0.31	0.26	0.39
		Trial number	-41.62	7.63	-5.46	-56.63	-26.56	28.09	1	**<0.001**				
	2	(Intercept)	64.70	2.99							0.05	0.26	0.24	
		Trial number	3.03	1.50	2.03	0.09	5.22	4.07	1	**0.044**				
		Time of day	2.90	1.42	2.04	0.10	6.73	4.12	1	**0.042**				
		Sex	8.03	4.81	1.67	-1.65	17.69	2.68	1	0.101				
ABD score	1	(Intercept)	425.00	106.70							0.04	0.04	0.00	0.15
		Trial number	-125.80	51.20	-2.46	-	-	6.05	1	**0.014**				
		Previous experience	-215.50	112.80	-1.91	-	-	3.68	1	**0.055**				
		Threat face location (L/R)*	162.30	100.80	1.61	-	-	2.60	1	0.107				
	2	(Intercept)	-49.50	91.47							0.05	0.10	0.02	

StimulusID1 *	50.58	127.75	0.40	-197.56	299.79	14.99	6	**0.020**
StimulusID2	141.13	127.84	1.10	-107.36	389.79			
StimulusID3	438.15	132.19	3.31	179.31	695.05			
StimulusID4	166.44	123.75	1.35	-74.45	406.91			
StimulusID5	12.49	131.22	0.10	-242.46	268.11			
StimulusID7	148.94	137.80	1.08	-119.32	416.72			

* Post-hoc exploration revealed a significant interaction of previous experience and threat face location (P<0.05).

* StimulusID6 was used as the reference category to aid interpretation of output

816

817 **Figure legends**

818 **Figure 1.** Two types of apparatus were developed and tested. 1a: The manually operated apparatus
819 with printed cards used to assess attention bias in Study 1 (manual). Top panel (A): the front of
820 apparatus, showing rectangular areas where stimuli were displayed and position of the camera,
821 including measurements in mm; Middle panel (B): the front of apparatus with filler (fruit) stimuli
822 revealed; Lower panel (C): apparatus from behind with sliding mechanism highlighted with a red
823 ring. 1b. The automated apparatus used for Study 2 (automated). Top panel (A): face (threat-neutral
824 conspecific face pair) stimuli; Middle panel (B): inter-trial interval; Lower panel: filler (fruit or
825 vegetable) stimuli.

826 **Figure 2.** Seven stimulus pairs were used in each of Study 1 and Study 2. In Study 1 stimuli were
827 presented printed on card. In Study 2 stimuli were presented as digitised images on screens. Stimuli
828 are shown ordered according to model intercepts obtained from Study 2 for ABD score, reading
829 across from top left to bottom right for greatest to smallest values (ID: mean ABD ± SE; #3: 365ms ±
830 117ms; #4: 99ms ± 78ms; #7: 82ms ± 101ms; #2: 76ms ± 102ms; #1: -11ms ± 72ms; #5: -51ms ±
831 74ms; #6: -65ms ± 90ms). Stimulus ID numbers recorded at the time of testing are shown in the top
832 right of each face pair for cross-referencing to the data set. Stimuli are available to download from
833 the folder 'Threat Neutral Face Pairs' in Witham & Bethell (2019).

834 **Figure 3.** Results for Study 1 (n= 322 data points from 66 females; mean ± SD). Dot sizes represent
835 the number of macaques represented at each data point. a) Duration of looking at the threat face
836 (THR) declined significantly over trials. b) Duration of looking at threat-neutral face pairs (TL)
837 declined significantly over trials. c) Attention bias difference scores (ABD) declined significantly over
838 trials (n=299 data points due to removal of three highly negative ABD outliers, one at trial #1 and
839 two at trial #3). d) Near-significant main effects of visual field and previous experience on ABD
840 subsequently reached significance when entered as an interaction term.

841 **Figure 4.** Results for Study 2 (n= 300 data points from 27 females and 16 males; mean ± SD). Dot
842 sizes represent the number of macaques represented at each data point. a) Duration of looking at
843 threat-neutral face pairs (TL) declined from trial 1 to trial 12. b) Duration of looking at threat-neutral
844 face pairs (TL) increased significantly between 09:00 and 15:00. c) Male macaques tended to have a
845 greater duration of looking at threat-neutral face pairs (TL) than did females. d) Attention bias
846 difference scores (ABDs) varied with stimulus ID with strong vigilance towards threat for stimulus #3,
847 mild vigilance towards threat for stimuli # 4, #7 and #2, and no vigilance or mild avoidance for
848 stimuli #1, #5 and #6.

849 **Figure 5.** Comparison of individual mean durations for looking towards the threat face for Study 1
850 and Study 2 (n=18).

851