How chimpanzees (*Pan troglodytes*) perform in a modified emotional Stroop task

Matthias Allritz (corresponding author)

Martin-Luther-Universität Halle-Wittenberg; Department for Differential Psychology and Psychological Assessment; Halle (Saale), GERMANY

Max Planck Institute for Evolutionary Anthropology; Department of Developmental and Comparative Psychology; Leipzig, GERMANY

Email: matthias_allritz@eva.mpg.de

Josep Call

University of St Andrews, School of Psychology and Neuroscience, UK

Max Planck Institute for Evolutionary Anthropology; Department of Developmental and Comparative Psychology; Leipzig, GERMANY

Peter Borkenau

Martin-Luther-Universität Halle-Wittenberg; Department for Differential Psychology and Psychological Assessment; Halle (Saale), GERMANY

This is an unedited manuscript that has not yet undergone copyediting, typesetting, and the author's review of proofs. It should therefore not be cited. For the published, citable version of this article, please refer to: Allritz, M., Call, J., & Borkenau, P. (2015). How chimpanzees (*Pan troglodytes*) perform in a modified emotional Stroop task. *Animal Cognition*. Advance online publication. doi: 10.1007/s10071-015-0944-3

Acknowledgements

The authors wish to thank the staff at Leipzig Zoo, particularly the zoo veterinarian and the chimpanzee caretakers for their various contributions to stimulus preparation and data collection. We thank Thurston Cleveland Hicks and Fabrizio Maffessoni for their contributions to stimulus preparation. We thank Alexander Weiss for providing a German version of the Hominoid Personality Questionnaire. We thank Daniel Geissler, Stefan Leideritz, Johannes Grossmann, and Sarah Peoples for providing personality ratings of the chimpanzees.

Compliance with ethical standards

Animal husbandry and research comply with the "EAZA Minimum Standards for the Accommodation and Care of Animals in Zoos and Aquaria", the "WAZA Ethical Guidelines for the Conduct of Research on Animals by Zoos and Aquariums" and the "Guidelines for the Treatment of Animals in Behavioral Research and Teaching" of the Association for the Study of Animal Behavior (ASAB).

ABSTRACT.

The emotional Stroop task is an experimental paradigm developed to study the relationship between emotion and cognition. Human participants required to identify the color of words typically respond more slowly to negative than to neutral words (emotional Stroop effect). Here we investigated whether chimpanzees (Pan troglodytes) would show a comparable effect. Using a touch screen, eight chimpanzees were trained to choose between two simultaneously presented stimuli based on color (two identical images with differently colored frames). In Experiment 1, the images within the color frames were shapes that were either of the same color as the surrounding frame, or of the alternative color. Subjects made fewer errors and responded faster when shapes were of the same color as the frame surrounding them than when they were not, evidencing that embedded images affected target selection. Experiment 2, a modified version of the emotional Stroop task, presented subjects with four different categories of novel images: three categories of pictures of humans (veterinarian, caretaker, stranger), and control stimuli showing a white square. Because visits by the veterinarian that include anaesthetization can be stressful for subjects, we expected impaired performance in trials presenting images of the veterinarian. For the first session, we found correct responses to be indeed slower in trials of this category. This effect was more pronounced for subjects whose last anaesthetization experience was more recent, indicating that emotional valence caused the slowdown. We propose our modified emotional Stroop task as a simple method to explore which emotional stimuli affect cognitive performance in nonhuman primates.

Keywords: chimpanzee, emotional Stroop, great apes, attentional bias, cognitive bias

Introduction

The study of attentional prioritization of stimuli of strong emotional valence has a long history in human cognitive science (e.g. MacLeod, Mathews, & Tata, 1986; Mathews & MacLeod, 1985). Numerous experimental paradigms have been developed to study how emotionally relevant stimuli are prioritized by visual attention (for reviews see Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, van Ijzendoorn, 2007; Mogg & Bradley, 2003; Yiend, 2010; Yiend, Barnicot, & Koster, 2013). Some of these paradigms require the human participant to make a manual response (such as pressing a button) to categorize a stimulus or stimulus feature, or to indicate the location of a stimulus. Additionally, the participant is presented with secondary stimuli or stimulus features which appear concurrently with or precede the task and which are irrelevant to it. Differences in responding (error rates and response latencies) as a function of the emotional valence of such secondary task features are typically interpreted as reflecting differential attentional prioritization of these features. In many cases, such effects of emotional valence are moderated by individual differences between participants, e.g. attentional prioritization of threatening stimuli might be restricted to, or of higher magnitude in, high or clinically anxious participants (Bar-Haim et al., 2007).

A classic example of such a paradigm is the emotional Stroop task (Mathews & MacLeod, 1985; for reviews see Bar-Haim et al., 2007; Phaf & Kan, 2007; Williams, Mathews, & MacLeod, 1996; Yiend, 2010). In the emotional Stroop task, human participants are typically required to name the colors of words that differ in emotional valence. Meta-analyses suggest moderate within-subject effects (i.e. longer latencies to respond to threatening or otherwise negative stimuli) in clinically anxious participants (Bar-Haim et al., 2007, Phaf & Kan, 2007).

Moderate effects could also be found for control participants, at least when stimuli of the same emotional valence were presented in blocks (Bar-Haim et al., 2007; McKenna & Sharma, 2004). A small number of studies used modified versions of the emotional Stroop task that used pictures (e.g. of human faces with different emotional expressions) instead of words as stimulus material, producing mixed results (Constantine, McNally, & Hornig, 2001; Kindt & Brosschot, 1997; Lavy & van den Hout, 1993; Mauer & Borkenau, 2007; Shibasaki, Isomura, & Masataka, 2014).

While the obvious adaptive value of attentional sensitivity (Lang, Davis, & Öhman, 2000; Öhman & Mineka, 2001) has inspired many studies that focus on stimuli that are deemed to be biologically relevant, such as "several types of vermin, facial expressions, but also blood and mutilations" (Phaf & Kan, 2007), it has also been suggested that stimuli which do not fall into this category of biologically prepared stimuli may acquire similar properties of enhanced attentional prioritization through learning (e.g. Öhman & Mineka, 2001; Yiend, 2010). In accordance with this, effects of attentional prioritization have been found for stimuli whose negative connotation is acquired rather than biologically prepared, such as taboo words (Mackay, Shafto, Taylor, Marian, Abrams, & Dyer, 2004; Siegrist, 1995) or pictures of weapons (e.g. Fox, Griggs, & Mouchlianitis, 2007). Moreover, stimuli that human participants have come to associate with negative outcomes as a result of aversive conditioning have been found to be attentionally prioritized in dot probe and visual search paradigms (Koster, Crombez, Van Damme, Verschuere, & De Houwer, 2004; Schmidt, Belopolsky, & Theeuwes, 2014). Finally, studies in the field of clinical psychology suggest that participants diagnosed with certain psychological disorders, such as posttraumatic stress disorder (Buckley, Blanchard, & Neill, 2000), or substance abuse (Field & Cox, 2008; Robbins & Ehrman, 2004) show consistent attentional prioritization of ontogenetically relevant stimuli associated with those disorders.

Paul, Harding, and Mendl (2005) put forward the idea that methods such as the visual dot-probe task or the emotional Stroop task that were originally developed to study cognitive biases in humans may be modified to study the link between emotion and cognition in nonhuman animals. In recent years, this idea has been put to the test in a few studies with nonhuman primates. Studies with rhesus macaques (*Macaca mulatta*) using the visual dot-probe paradigm have revealed in this species an attentional bias for aggressive facial expressions of conspecifics (King, Kurdziel, Meier, Lacreuse, 2012; Lacreuse, Schatz, Strazullo, King, & Ready, 2013), but no attentional bias for neutral faces of newborn (rather than adult) conspecifics (Koda, Sato, & Kato, 2013). Shibasaki and Kawai (2009) demonstrated an attentional prioritization of pictures of snakes over pictures of flowers in Japanese macaques (*Macaca fuscata*) in a study using the visual search paradigm. In another study using the visual search paradigm, Marzouki, Gullstrand, Goujon, and Fagot (2014) found that baboons (*Papio papio*) located a T-shaped target among L-shaped distractors more slowly in trials that followed the spontaneous expression of negatively, rather than neutrally or positively, valenced behaviors by the subjects. Finally, several studies using the cognitive judgment bias paradigm have investigated the effects of emotions on decision making in rhesus macaques (*Macaca mulatta*; Bethell, Holmes, MacLarnon, & Semple, 2012), tufted capuchins (*Cebus apella*; Pomerantz, Terkel, Suomi, & Paukner, 2012) and chimpanzees (*Pan troglodytes*; Bateson & Nettle, 2015), as well as many nonprimate species (for a review see Bethell, 2015). However, to our knowledge no experiment has yet applied the emotional Stroop task or variations thereof to study the relationship between emotion and cognition in nonhuman primates.

The aim of this study was twofold: first, we intended to develop a novel, simple experimental paradigm suitable for chimpanzees and other nonhuman primates by building on a

modified pictorial version of the emotional Stroop task introduced by Mauer and Borkenau (2007). Our second aim was to investigate whether the emotional valence of pictures presented concurrently with this color discrimination task would indeed affect the performance of the subjects. We chose pictures as stimuli, because experimental studies with chimpanzees have shown that chimpanzees are affected by the emotional valence of pictorial or video content (see Bovet & Vauclair, 2000, for a review of picture recognition in nonhuman animals). Chimpanzees have been shown to exhibit accelerated heart rates in response to viewing photographs of an aggressive conspecific (Boysen & Berntson, 1989), changes in peripheral skin temperature when viewing video scenes of negative emotional valence (Parr, 2001), enhanced recognition of pictures of aggressive (rather than neutral) conspecific interactions (Kano, Tanaka, & Tomonaga, 2008), and differential event-related brain potentials in response to viewing affective (rather than neutral) pictures (Hirata et al., 2013).

The chimpanzee subjects in this study were presented with a simple discrimination task in which the subjects needed to select one of two stimuli presented simultaneously on a touch screen. For each trial, two identical pictures that only differed in the color of a frame surrounding them served as stimuli. Each subject was trained to always select the same color on every trial. The first experiment was designed to establish that, in spite of being trained to respond solely based on stimulus frame color, subjects' performance would nonetheless be affected by the pictorial content embedded in those color frames. Therefore, non-social abstract stimuli (geometric shapes) with color features relevant to the discrimination task were used to examine whether these features would impair or improve performance in predicted directions. In the second experiment we presented subjects with stimuli that differed in their (presumed) ontogenetically acquired emotional valence (pictures of human beings that had different

relationships with the chimpanzee subjects) to investigate whether these would also affect performance in predicted directions. Finally, we collected trait ratings from animal caretakers to explore whether individual differences in personality might moderate the effects of emotional valence.

Training Procedure

Method

Subjects

All subjects participating in this study were from the same chimpanzee group housed at Wolfgang Köhler Primate Research Center (WKPRC) in Leipzig, Germany, which included 6 male and 12 female chimpanzees (age ranging from 3 to 37 years) at the beginning of this study. All of the subjects participating in this study had been successfully trained to use the touch screen setup before color discrimination training began. Four male and seven female chimpanzees participated in the training phase of this study. One adult female was excluded over the course of training because exploration of the experimental setup by her dependent offspring made individual testing impossible. This resulted in a final sample of four male and six female chimpanzees (mean age in years M =20.80, SD = 13.72) who completed the training phase of the study. All great apes at Leipzig zoo are housed in groups with regular access to large indoor and outdoor enclosures. Subjects also have access to sleeping and observation rooms in which non-invasive experimental studies are conducted. Subjects receive a regular diet of fruit, vegetables and animal food and they are never deprived of food or water.

Apparatus

All tests were conducted in the chimpanzee observation rooms at WKPRC. For the experimental tasks we used a custom-made setup. Outside the testing cage, the experimenter set up a computer that was connected to two monitors as well as two audio speakers which provided auditory feedback to the subjects' performance and which were located in front of the testing cage. Subjects operated a transparent optical touch screen (Nexio NIB-190B infrared

touchscreen, 19 inches in diameter) embedded into a robust metal panel that was part of the cage mesh. Behind this see-through touch screen, one monitor (ViewSonic VG930m, 19 inches, resolution of 1280 x 1024 pixels, frequency of 60 Hz) was mounted to display the experimental stimuli to the subjects. The touch screen was connected to the experimenter's computer via USB cable and was calibrated using the iNexio Touch Driver software so that spatial positions touched on the touch screen would correspond to the same spatial positions on the monitor mounted behind it. A second monitor enabled the experimenter to follow the experiment's progress. All experimental procedures including stimulus presentation and response collection were carried out using E-Prime 2.0.8.90 running under Windows 7.

Stimuli

All stimuli covered an area of 350 × 350 pixels (ca. 10.31 cm x 10.31 cm) and consisted of an image that was 300 × 300 pixels in size (ca. 8.83 cm x 8.83 cm) which was surrounded by a frame with a width of 25 pixels (ca. 0.74 cm) that was either blue (RGB 0,0,255) or yellow (RGB 255,255,0). The images within this color frame consisted of photographs (in training conditions and in Experiment 2) or geometric shapes (in Experiment 1) that were presented in front of a white background. All stimuli were prepared using Adobe Photoshop CS2 and CS6.

For the color discrimination training, 50 images of random human artifacts presented on a white background were used. For the color discrimination transfer test (see below), 50 new images of human artifacts were used. Pictures of human artifacts used for the training and transfer test stimuli included pictures of clothes and accessories, cutlery and tableware, furniture, household appliances, musical instruments, sports equipment, technical and electronic

equipment, tools, toys, and vehicles. The pictures used in these training conditions were assembled using Google Images search.

General Procedure

All stimuli were presented on a black (RGB 0,0,0) background. Every trial was initiated by the subject by touching a white start key located in the center of the screen. This was followed by a 500 ms delay upon which the target (e.g. an image with a yellow frame) and the distractor (e.g. the same image with a blue frame) appeared in locations of equal horizontal distance to the start key (distance between center of screen and center of stimulus 320 pixels (ca. 9.42 cm), see Fig. 1a). If the subject selected the target, the stimuli disappeared, a high-pitched chime was played and the subject was rewarded by the experimenter with a piece of food. After an intertrial interval of 1500 ms the start key for the next trial was presented. If the subject selected the distractor, a low-pitched tone was played, the subject was not rewarded and the intertrial interval was extended by an additional 3000 ms time out, resulting in a 4500 ms intertrial interval before the next start key appeared.[1] This trial procedure was the same for the color discrimination training A (see below), the color discrimination transfer test, and the experimental conditions.

For correct choices, subjects were rewarded with pieces of apple. In some sessions, two of the subjects were rewarded with half a grape or a banana pellet on every fifth correct trial to ensure continuous participation. Occasionally sessions were terminated prematurely because (a) the subject stayed inactive for more than five minutes, or (b) the subject showed clear signs of aggression (e.g. hitting the screen). These sessions were then continued on the next testing day. The same rules for premature termination also applied to the experimental phases of this study (both the refresher test sessions as well as the experimental sessions).

Target frame color (i.e. whether the blue or the yellow frame stimuli constituted the target) was counterbalanced across subjects with blue being the target frame color for five of the original eleven subjects. In the final sample of eight (Experiment 1) or seven (Experiment 2) subjects, blue was the target frame color for three subjects.

Color Discrimination Training A

During training, subjects completed 100 trials on each testing day. In each trial the subject was presented with a target (one of the 50 images surrounded by e.g. a blue color frame) and a distractor (the same image surrounded by a yellow color frame). Each target-distractor-combination was presented twice in each session, once with the target on the right side of the screen and once with the target on the left side. The resulting 100 trials were completed by the subject in a randomized order, with the sole restriction that target stimuli were not presented on the same side of the screen in more than two consecutive trials. Once the subject's performance exceeded 80 correct trials in each of two consecutive sessions, the subject proceeded to the color discrimination transfer test. If the subject failed to reach this criterion within 40 sessions, it proceeded to the color discrimination training B instead.

Color Discrimination Training B

Four subjects failed to reach criterion within 40 sessions of color discrimination training A. These subjects received additional training with a modified trial procedure that was designed to reduce side and perseveration biases and to maximize learning from feedback. The stimuli were the same that were used in color discrimination training A. The modified trial procedure was as follows. Subjects initiated each trial by pressing a white start key located in the center of

the screen. This was followed by a 500 ms delay upon which target and distractor appeared in two out of eight possible locations (using every location except the central location in a virtual 3 x 3 grid on the screen). If the subject selected the distractor, the target disappeared and the distractor remained on screen for an additional 500 ms. This was accompanied by a low-pitched tone indicating no reward. After an additional 1000 ms of blank screen, the presentation of both stimuli was repeated with target and distractor appearing in the same locations as before. If the subject selected the target, the distractor disappeared and the target remained on screen for an additional 500 ms. This was accompanied by a high-pitched chime and the subject was rewarded by the experimenter with a piece of food. After an intertrial interval of 500 ms the start key for the next trial was presented.[1] There was no restriction to the number of repetitions the subject had to complete, i.e. subjects received repetitions of the same trial until they selected the target. Target and distractor positions were randomly determined before trial onset but remained the same for each trial repetition. Each of the 50 target-distractor pairs was presented in two trials per session, resulting in 100 trials in total per session. Once the subject's performance exceeded 80 trials with correct first choice on each of two consecutive sessions, the subject returned to color discrimination training A (see Table 1). If the subject failed to reach this criterion within 40 sessions, it was dropped from the study.

Color Discrimination Transfer Test

To rule out the unlikely possibility that subjects had learnt to respond correctly separately for each individual stimulus pair over the course of training rather than acquiring a generalized rule based on stimulus frame color, a transfer test was presented to subjects upon reaching criterion in color discrimination training A. Trial procedure and performance criterion in this

transfer test were identical to color discrimination training A, except for the fact that 50 completely new images were used as stimuli embedded in the color frames.

- insert Figures 1a, 1b, and 1c around here -

Results and Discussion

Table 1 illustrates how many sessions each subject completed before reaching criterion (more than 80% correct responses in two consecutive sessions) for each training condition. As can be seen, four subjects did not reach criterion within forty sessions of color discrimination training A and thus received additional training sessions of color discrimination training B until reaching criterion (fourth column). Two of these four subjects reached criterion in training B and subsequently reached criterion after additional sessions of training A (fifth column). All eight subjects who eventually reached criterion in training A proceeded to the color discrimination transfer test. As can be seen in the last column, all eight subjects reached this criterion considerably faster than in training phase A, evidencing the acquisition of a generalized rule based on stimulus frame color. By successfully reaching criterion in the transfer test, all of these subjects qualified for the experimental studies.

- insert Table 1 around here -

Experiment 1 – Color interference task

As described in the introduction, the aim of Experiment 1 was to determine whether subjects' performance in the task (selecting the correct stimulus based only on the color of its

frame) would be affected by the task-irrelevant pictorial content embedded in those color frames. In order to create an experimental situation that would maximize the probability of giving rise to such effects of embedded content on task performance, we presented subjects with novel target and distractor stimuli in which the embedded content consisted of geometrical shapes (see Fig. 1b) that were identical in shape but differed in their color (blue or yellow). In *congruent* trials, the geometric shapes were of the same color as the frames surrounding them, in *incongruent* trials these geometric shapes were of the alternative color (e.g. such that a yellow shape would be embedded in the blue color frame. We predicted that subjects would show lower accuracy in incongruent trials as opposed to congruent trials. We also predicted that within correct trials, subjects would exhibit longer latencies in incongruent trials than in congruent trials. Such an interference effect of embedded picture content on task performance, if it existed, could be regarded as evidence that subjects are indeed affected by content that is objectively irrelevant to making a correct selection. The apparatus and the trial procedure were identical to the training phase.

Method

Subjects

All eight subjects who had successfully completed color discrimination training participated in Experiment 1. This resulted in a sample of 3 males and 5 females (mean age of all subjects in years at the beginning of this experiment: $M = 17.75$, $SD = 12.30$). It should be noted that before participating in Experiment 1 these eight subjects participated in an additional experiment utilizing this paradigm comprising four to seven sessions in total that could not be

312 considered for this study due to technical difficulties during data collection for several subjects.
313 However, all of the stimuli used in Experiments 1 and 2 were previously unfamiliar to the
314 subjects, except where noted.
315
316 *Stimuli*
317 For the color interference task, stimuli contained images of four different geometrical
318 shapes (square, circle, flower, star) that were either blue or yellow and presented on a white
319 background, with a frame surrounding this image which was either of the same color (congruent
320 condition) or of different color (incongruent condition). Additionally, five black-and-white
321 photographs of everyday objects (book, mug, pencil sharpener, plate, watering can) presented on
322 a white background were used as control stimuli (see Fig. 1b for example stimuli used in the
323 color interference task). These control stimuli, with which the subjects were already familiar
324 from a previous experiment, were selected to ensure minimal interference with color
325 discrimination performance.
326
327 *Design*
328 One to four days before the experiment, subjects were required to pass a refresher test
329 that consisted of one session of the color discrimination transfer test. If subjects' performance
330 exceeded 70 % on this refresher test (a criterion which all subjects met on first attempt), they
331 began participating in the experiment on the next testing day. This more relaxed criterion was
332 chosen to avoid overtraining and thus ceiling effects in subjects' accuracy across conditions. The
333 experiment consisted of two sessions, each presenting subjects with a total of 120 trials that
334 included 80 test trials (congruent and incongruent trials) and 40 control trials. Each test trial

presented one of four different geometrical shapes in the center of both target and distractor color frame. The shapes were either of the same color as the color frames surrounding them (congruent trials) or of the alternative color (incongruent trials). Targets could either appear on the left or the right side of the screen. These parameters combined to 4 (shape) x 2 (congruence) x 2 (target side) = 16 unique test trial configurations. Each of these 16 unique test trial configurations was presented 5 times over the course of a session, yielding 80 test trials (40 congruent and 40 incongruent trials). Each control trial presented one of five different black and white photographs, which were familiar to the subjects, in the center of both target and distractor color frame. Again, targets could either appear on the left or the right side of the screen, yielding 5 (photo) x 2 (target side) = 10 unique control trial configurations, of which each was presented four times per session, resulting in a total of 40 control trials. In both experimental sessions all test and control trials were presented in random order with the sole restriction that target stimuli would not be presented on the same side of the screen for more than two consecutive trials. For one subject Session 1 and Session 2 each had to be split into two parts (conducted on different testing days) because the subject stayed inactive for more than five minutes during the course of the session.

Data analysis

In order to compare subjects' performance across conditions, the mean accuracy (percentage of correct trials across both sessions) was calculated for each subject for each condition. We performed a one-way repeated measures ANOVA with condition (levels: control, congruent, incongruent) as factor and mean accuracy as dependent variable. As mean accuracy

represents a proportion, the data was arcsine-transformed to approximate normality before further analysis (Cohen & Cohen, 1983).

In order to compare subjects' response latencies across sessions the median response time for correct trials was calculated, again for each subject for each condition across sessions. These individual response latency scores were then subjected to a one-way repeated measures ANOVA with category (levels: control, congruent, incongruent) as factor, and latency medians as dependent variable. Degrees of freedom were Greenhouse-Geisser-corrected for all analyses. All statistical tests were two-tailed.

Inspection of video recordings of all sessions revealed that in a small number of trials problems with response recording occurred, i.e. at least one touch to either stimulus was not immediately followed by appropriate program feedback. Such problems occurred in 24 of 1908 trials (the 12 remaining trials could not be evaluated because a subject was blocking the view), which corresponds to 1.26 % of trials. Further inspection suggested that these instances could almost entirely be attributed to the manner in which the infrared touchscreen was operated by subjects in these trials (e.g. during the non-registered touch, one of the subject's fingers was touching the background area, thereby blocking the touch screen program temporarily from recording further input). All 24 trials were excluded from analysis. Including these trials in data analysis did not affect results substantially.

Results

Accuracy

Figure 2a depicts mean accuracy scores for the different conditions. There was a significant effect of condition on accuracy, $F(1.93, 13.54) = 19.44$, $p < .001$. Pairwise

comparisons revealed that chimpanzees performed significantly worse in incongruent trials than they did in congruent trials, $t(7) = -6.34$, $p < .001$, or in control trials, $t(7) = -4.33$, $p = .003$, whereas there was no significant difference between congruent and control trials, $t(7) = -1.23$, $p = .258$.

Latency

Figure 2b depicts mean latency scores for the different conditions. There was a significant effect of condition, $F(1.38, 9.65) = 6.90$, $p = .020$. Paired samples *t*-tests revealed that chimpanzees responded significantly faster in congruent trials than in incongruent trials, $t(7) = -2.95$, $p = .022$, or control trials, $t(7) = -4.13$, $p = .004$, but there was no significant difference between response latencies in incongruent vs. control trials, $t(7) = 1.06$, $p = .326$.

- insert Figures 2a and 2b around here -

Discussion

Subjects made more errors in incongruent trials than in congruent or control trials. Considering correct trials only, subjects were faster to complete congruent trials than incongruent or control trials. These findings are in accordance with our hypothesis that in spite of being trained to ignore pictorial content and respond based on frame color only, subjects' performance was indeed affected by the pictorial content embedded in the color frames. The results of the first experiment may thus be regarded as a "proof of concept", evidencing that under certain conditions frame content may affect the accuracy and speed of frame color discrimination. The second experiment was designed to investigate whether this effect could also be detected for

stimuli that differed primarily in terms of their (presumed) emotional relevance to subjects – that is whether subjects would exhibit an effect resembling the emotional Stroop effect.

Experiment 2 – Modified emotional Stroop task

In Experiment 2, we presented subjects with color photographs of human beings embedded in the color frames, and with control stimuli in which the color frame contained only a white square. The color photographs belonged to three categories based on the relationships that the depicted humans had with the chimpanzee subjects (veterinarian, caretakers, unfamiliar humans). While it is in the interest of all the staff at Leipzig Zoo to maintain and further animal welfare and well-being, stressful encounters as part of medical procedures cannot always be avoided. In particular, visits by the zoo veterinarian that include anaesthetization are stressful to most chimpanzee subjects. We thus expected the emotional valence associated with photographs depicting the veterinarian to be negative for all subjects who had had at least one anaesthetization experience before Experiment 2 was conducted. Based on the human literature on attention to emotional stimuli, we expected interference effects (impaired performance in the color discrimination task) to be most pronounced for these (presumably negative) stimuli, that is, we predicted lower accuracy as well as longer latencies in correct trials for stimuli depicting the veterinarian than for any other stimulus category (caretaker, unfamiliar humans, control). For the caretakers and unfamiliar humans, it is more difficult to hypothesize which emotional reaction a particular picture might evoke in a particular subject. We thus had no hypotheses with regard to differences in accuracy or latency between these stimulus categories. Because the number of unique stimuli used in this experiment was quite small (four stimuli per category), we also

examined whether interference effects for negative stimuli may be subject to habituation, that is whether they would decrease over sessions.

Because we assume the negative valence of photographs of the veterinarian to be ontogenetically acquired, individual experience with anaesthetization has to be taken into account. The time since the last anaesthetization experience differed considerably for the subjects participating in this experiment, with one subject having never had an anaesthetization. Consequently, we expected the interference effects for stimuli from the veterinarian category to be stronger for those subjects whose experience with the anaesthetization procedure was more recent, and we expected weaker effects for the subject who had not had any anaesthetization experience yet (but who had also been visited by the veterinarian before).

In humans, interference effects of negative stimuli in the emotional Stroop task are often moderated by individual differences in personality (e.g. Bar-Haim et al., 2007; Mauer & Borkenau, 2007). Therefore, we also computed anxiety and aggression scores which were derived from trait ratings provided by human raters who were familiar with the chimpanzees, to investigate whether interference effects associated with negative stimuli might be more pronounced in chimpanzees that were described as more anxious or, alternatively, more aggressive by human raters. The apparatus and the trial procedure were identical to the training phase.

Method

Subjects

Seven subjects participated in Experiment 2. One female chimpanzee that had previously participated in Experiment 1 could not participate in Experiment 2 because she avoided

operating the touch screen in multiple attempts of conducting the refresher test. This exclusion yielded a sample of 3 males and 4 females for Experiment 2 (mean age of all subjects in years at the beginning of this experiment: $M = 17.29$, $SD = 13.21$).

Stimuli

Three different stimulus categories including pictures of humans were used, each comprising four different images (see Figure 1c and Table 2 for details). The category "stranger" included two photographs of each of two different humans unfamiliar to the subjects, one image of each stranger showing the face only, and the other showing the actor from the waist up, holding an object (in this case a backpack) in front of them. The category "caretaker" included photographs of two different caretakers whom the subjects see and interact with regularly. Both caretakers had known each subject participating in the study for at least eight years. The category included one image of each caretaker showing the face only and one image of each caretaker showing the actor from the waist up, holding an object (in this case a food bucket without visible food) in front of them. The category "vet" included four pictures of the zoo veterinarian, two images of the veterinarian showing the face only (one with and one without work gear typically worn when encountering the subjects) and two images of the veterinarian showing the actor from the waist up (again, one with and one without work gear), holding a blowpipe, typically used to anaesthetize animal subjects, in front of his face, aiming at the viewer. All human actors were male. One additional stimulus containing only a blank white square embedded in the color frame was used as a control stimulus. To maximize recognizability and ecological validity, we presented subjects with color, rather than black and white images. Photoshop CS6 was used to match stimuli as best as possible for their *luminosity* parameters across stimulus categories

(stranger, caretaker, vet) and image types (face image, upper body image). For a list of all stimuli used in Experiment 2, see Table 2.

- insert Table 2 around here -

Design

One to four days before the experiment, subjects were required to pass a refresher test that consisted of one session of the color discrimination transfer test (see Experiment 1). The performance of all subjects exceeded 70 % in this refresher test and they began participating in the experiment on the next testing day.

The experiment consisted of three sessions. Within each session, we presented subjects successively with small test blocks that included four stimuli from the same category, followed by one control trial. We arranged stimuli in this order because studies with humans have shown emotional Stroop effects to be most pronounced when stimuli of the same valence category are presented in blocks (Bar-Haim et al., 2007; McKenna & Sharma, 2004). Hence, in this study stimuli from the same valence category were also presented in blocks. However, frequent repetitions of the same stimuli often result in habituation in studies using the emotional Stroop task (e.g. Ben-Haim, Mama, Icht, & Algom, 2014; Witthöft, Rist, & Bailer, 2008). Because in this study we used only four unique stimuli of each category, we attempted to minimize possible within-block habituation effects by reducing the number of trials within blocks to four. Finally, each block of stimuli from the same category was followed by one control trial, thus separating blocks of stimuli from different categories. Control trials were interspersed in this manner to minimize carry-over effects (stimulus valence affecting performance in subsequent trials) that

have been reported to occur in emotional Stroop tasks (Algom, Chajut, & Lev, 2004; Frings, Englert, Wentura, & Bermeitinger, 2010; McKenna & Sharma, 2004; Waters, Sayette, & Wertz, 2003).

In each of the three sessions, subjects completed a total of 125 trials, including 29 control trials, 32 stranger trials, 32 caretaker trials, and 32 vet trials. Whether the target would appear on the left or on the right side was randomly determined for each trial. Each session began with a warm-up block of five control trials which was followed by 24 test blocks, with each test block consisting of five trials in total: the first four trials presented the subject with all four unique stimuli from the same category (stranger, caretaker, or vet), while the fifth trial was a control trial. Within the four test trials of each test block, the order of stimuli presented was counterbalanced such that across the three sessions, every subject was presented with all possible orders of the four stimuli from that category exactly once. The 24 test blocks of each session were further organized in segments, with each segment consisting of three test blocks (one from each category in counterbalanced order). Thus, each session (excluding the five warm-up trials at the beginning) consisted of a succession of eight segments. Consequently, subjects were presented with eight test blocks of each category per session. For two subjects Session 1 had to be split into two parts (conducted on different testing days) because the subjects exhibited clear signs of aggression during the first testing session (see Results section).

Personality trait ratings

In order to obtain personality measures, four raters filled out a German version of the Hominoid Personality Questionnaire (HPQ; King & Figueredo, 1997; Weiss et al., 2009) for all 17 chimpanzees (6 males and 11 females, mean age $M = 22.06$, $SD = 12.92$) that were at the time

of data collection part of the same housing group as the eight subjects who participated in the experimental studies. Two of the four raters were animal caretakers and two raters were research assistants who frequently carry out behavioral observations on all subjects from that group. Each rater had at least 1.5 years of experience with each subject. The current version of the HPQ consists of 54 items (e.g. anxious, friendly, intelligent) that are complemented by behavioral descriptions (e.g. "ANXIOUS: Subject often seems distressed, troubled, or is in a state of uncertainty"). The rater indicates on a Likert scale that ranges from 1 ("Displays either total absence or negligible amounts of the trait.") to 7 ("Displays extremely large amounts of the trait.") to which extent he or she finds the trait to be characteristic of the subject in question. Trait ratings were provided by all four raters for all 17 subjects for all 54 items. Only a subset of items was considered for further analysis in the context of this study because of the item's obvious relevance (face validity) to the personality domain of anxiety (anxious, cautious, excitable, fearful, timid) or aggression (aggressive, bullying, irritable; and reverse coded: affectionate, friendly, gentle, helpful, sympathetic).

Data analyses

In order to compare subjects' performance across conditions, the mean accuracy (percentage of correct trials) was calculated for each subject for each condition in each session. We performed a two-way repeated measures ANOVA with session and condition (levels: control, stranger, caretaker, vet) as factors and mean accuracy as dependent variable. Again, accuracy data was arcsine-transformed to approximate normality.

In order to compare subjects' response latencies, the median response time for correct trials was calculated for each subject for each condition in each session.[2] These individual

response latency scores were then subjected to a two-way repeated measures ANOVA with session and condition (levels: control, stranger, caretaker, vet) as factors, and the individual latency medians as dependent variable. In order to examine whether interference effects would decrease across sessions (as a result of habituation) we also analyzed the data for a possible interaction between condition and session. Degrees of freedom were Greenhouse-Geisser-corrected for all analyses.

To quantify individual differences in task interference elicited by the presence of negative stimuli, we computed individual interference scores, as is frequently done in emotional Stroop paradigms. Because at the group level subjects showed habituation to the veterinarian stimuli over the course of the three sessions (see results section and Fig. 3b), we restricted the analysis of individual differences to Session 1. Interference scores were computed as the differences between response time in (correct) trials of the veterinarian condition and each of the other conditions, yielding three interference scores for each subject. These interference scores quantify for each subject to what extent the veterinarian stimuli (as opposed to other stimuli) interfere with and thus slow down the subject's performance. Pearson correlation coefficients were computed to investigate the relationship between these interference scores and time passed since the last anaesthetization, as well as between interference scores and personality scores. All statistical tests were two-tailed.

As discussed for Experiment 1, problems with response recording occurred in a small number of trials (51 of 2624 evaluated trials, i.e. 1.94 %). Again, all of these trials were excluded from analysis. Including these trials in data analysis did not affect results substantially.

Results

563

564 *Accuracy*

565 Figure 3a shows performance in the different conditions across sessions for those six
566 subjects who had had experienced anaesthetization. The Session x Condition ANOVA revealed a
567 significant main effect of condition, $F(1.90, 9.51) = 11.30$, $p = .003$, as well as a significant main
568 effect of session, $F(1.87,9.37) = 7.45$, $p = .012$, but no significant interaction, $F(2.73, 13.66) =$
569 1.86, $p = .187$. Pairwise comparisons of accuracy (across all three sessions) between the vet
570 condition and the other conditions (*t*-tests for paired samples) revealed that chimpanzees
571 performed worse in veterinarian than in control trials, $t(5) = -5.51$, $p = .003$, whereas no
572 significant difference was found between vet trials and stranger trials, $t(5) = -.68$, $p = .524$, or
573 between vet trials and caretaker trials, $t(5) = -1.96$, $p = .107$.

574 Following a suggestion by an anonymous reviewer we investigated whether the presence
575 vs. absence of the gear that the veterinarian typically wears when anaesthetizing subjects (see
576 Stimuli section) had an effect on the subjects' performance. To this end, we conducted a number
577 of analyses that were restricted to data from veterinarian trials. Similar to the main analyses of an
578 effect of condition described above, we analyzed whether there was an effect on accuracy by
579 conducting a two-way repeated measures ANOVA with session and condition (levels: vet with
580 work gear, vet without work gear) as factors and arcsine-transformed mean accuracy as
581 dependent variable. This analyses did not reveal a significant effect of gear presence, $F(1.00,$
582 $5.00) = .48$, $p = .518$, or session, $F(1.64, 8.18) = 1.42$, $p = .288$, nor a significant interaction of
583 the two factors, $F(1.42, 7.12) = .60$, $p = .519$. An analysis restricted to Session 1 (t-test for paired
584 samples) did not reveal a difference in accuracy between trials with work gear present vs. absent
585 that reached conventional levels of statistical significance, $t(5) = 2.16$, $p = .083$.

Latency

Figure 3b shows latency in correct trials in the different conditions across sessions for all 6 subjects who had had experienced anaesthetization in the past. The Session x Condition ANOVA revealed a significant main effect of condition, $F(1.48, 7.38) = 15.08$, $p = .003$. The main effect of session was marginally significant, $F(1.87,9.36) = 3.40$, $p = .080$, as was the interaction between the two factors, $F(2.16, 10.80) = 3.65$, $p = .059$. Because the presence of an interaction makes it difficult to interpret main effects (Underwood, 1997), we further investigated this interaction by analyzing the data for all three sessions separately. One-way ANOVAs revealed significant effects of condition in Session 1, $F(1.32,6.60) = 11.02$, $p = .011$, Session 2, $F(1.36,6.81) = 6.38$, $p = .034$, and Session 3, $F(1.24,6.20) = 7.06$, $p = .033$. Pairwise comparisons (paired samples t-tests) revealed that in Session 1 chimpanzees responded more slowly in trials presenting vet stimuli than in all other conditions (control: $t(5) = 3.59$, $p = .016$, caretaker: $t(5) = 2.67$, $p = .044$, stranger: $t(5) = 3.65$, $p = .015$). In Session 2, responses in trials presenting vet stimuli were significantly slower only in comparison to control stimuli, $t(5) = 5.89$, $p = .002$, but not caretaker, $t(5) = 1.75$, $p = .140$, or stranger stimuli, $t(5) = 1.06$, $p = .337$. In Session 3, responses in trials presenting vet stimuli were significantly slower both in comparison to control stimuli, $t(5) = 2.78$, $p = .039$, and stranger stimuli, $t(5) = 3.09$, $p = .027$, but not in comparison to caretaker stimuli, $t(5) = 1.74$, $p = .142$.

As described in the Accuracy section, we also explored whether the presence vs. absence of work gear in the veterinarian trials had an effect on response latency in correct trials. We conducted an ANOVA with session and conditions as factors and median response latencies as dependent variable. Neither the effect of session ($F(1.04, 5.20) = 3.78$, $p = .107$), nor condition

($F(1.00, 5.00) = 3.70$, $p = .112$), nor the interaction ($F(1.04, 5.18) = .96$, $p = .375$) reached conventional levels of statistical significance. Considering data from Session 1 alone, in spite of a sizable difference in response latency between the two conditions (mean latency when gear was present: $M = 1320.67$ ms, when gear was absent: $M = 1024.75$ ms), the effect was not statistically significant, $t(5) = 1.27$, $p = .259$.

- insert Figures 3a and 3b around here -

Performance and anaesthetization experience

Except for one male chimpanzee, all subjects had had at least one anaesthetization experience when the study was conducted. The time since the last anaesthetization ranged from 184 to 2676 days (ca. 6 to 88 months, $M = 40.15$, $SD = 30.05$). Figure 3c shows interference scores (differences in response latency between vet stimuli and each of the other stimulus categories) as a function of time since the last anaesthetization. Among the six subjects who had had anaesthetization experience, time passed since the last anaesthetization correlated strongly with interference scores from the first session. These correlations were significant for interference scores based on control stimuli, $r(4) = -.92$, $p = .009$, and for interference scores based on stranger stimuli, $r(4) = -.88$, $p = .020$, whereas the correlation between time since anaesthetization and interference scores based on caretaker stimuli was marginally significant, $r(4) = -.81$, $p = .051$. In addition to the statistical evidence, it should be noted that two subjects whose latest anaesthetization had been fairly recent in comparison to other subjects (6 months and 35 months) exhibited noticeable emotional reactions in the presence of vet stimuli during their first session, including backing away from the touch screen, vocalizations, ignoring food

rewards in spite of continued participation (one subject), hitting and/or kicking the touch screen, and even breaking it (one subject). As mentioned above, these sessions were terminated prematurely and continued on the next testing day (without any further emotional reactions of this magnitude). Finally, the subject that did not have any prior anaesthetization experience did not exhibit response latencies that were substantially slower in the veterinarian condition than in the other two conditions that included pictures of humans (interference score based on control stimuli: 69.5 ms; based on stranger stimuli: 13.5 ms; based on caretaker stimuli: -9 ms).

- insert Figure 3c around here -

Performance and personality

Interrater reliability was determined for each item by calculating ICC(3, 4) for all of the 13 relevant items from the Hominoid Personality Questionnaire. Only items with reliability values higher than .5 were considered for further analysis, which led to the exclusion of the items "excitable" and "affectionate". Interrater reliabilities for the remaining 11 items ranged from .53 to .74, with an average of .65. Mean ratings (across raters) for these 11 items were subjected to further analysis. Cronbach's α was determined both for the scale comprising the remaining four items indicating anxiety (anxious, cautious, fearful, timid), as well as for the scale comprising the remaining seven items indicating aggression (aggressive, bullying, irritable; and reverse coded: friendly, gentle, helpful, sympathetic), revealing excellent internal consistency for the anxiety scale ($\alpha = .95$) and for the aggression scale ($\alpha = .91$). Consequently, anxiety scores (the mean of the four anxiety items) and aggression scores (the mean of the seven aggression items) were computed for every subject who participated in Experiment 2. Among the six subjects who had

had anaesthetization experience, interference scores for the first session (latency difference between vet stimuli and other stimuli) did not correlate significantly with anxiety scores (interference scores based on control stimuli: $r(4) = -.02$, $p = .973$; caretaker stimuli: $r(4) = -.11$, $p = .843$; stranger stimuli: $r(4) = .12$, $p = .816$), nor did they correlate significantly with the aggression scores (control stimuli: $r(4) = .22$, $p = .674$; caretaker stimuli: $r(4) = .41$, $p = .416$; stranger stimuli: $r(4) = .42$, $p = .402$).

Discussion

The purpose of Experiment 2 was to test the hypothesis that stimuli of negative emotional valence (pictures of the zoo veterinarian) would interfere with the performance of chimpanzee subjects in a color discrimination task (resulting in lower accuracy and slower responding). Our prediction with regard to response latency was confirmed by the data: breaking down the interaction between session and condition revealed that in the first session (when subjects saw all stimuli for the first time), response latencies on correct trials were slower in trials presenting vet stimuli than for any other stimulus class. Slow-down effects of this magnitude for all other stimulus classes were not observed in subsequent sessions. This difference between the first session and later sessions appears likely to be a result of habituation – at the beginning of Session 2 each subject had already seen every stimulus (including the four veterinarian stimuli) eight times. Additionally, for Session 1, we found the slowing of responses in trials presenting the veterinarian stimuli (in comparison to other stimulus categories) to be more pronounced in subjects for whom less time had passed since the last anaesthetization procedure. Thus, it appears plausible that increased task interference was indeed a result of negative emotional valence associated with these stimuli.

With regard to accuracy, while there was a main effect of condition across sessions, pairwise comparisons revealed a significant difference only between trials presenting veterinarian stimuli and control trials, but not between veterinarian stimuli and the other two human picture categories. These weaker effects of stimulus valence on accuracy mirror results in emotional Stroop tasks with human participants. In humans accuracy is typically at ceiling in all valence conditions and interference effects are manifested only in response time differences between conditions. Our chimpanzee subjects had had extensive training with the task, resulting in good to very good average performance across the different conditions. Additionally, while the negative stimuli used in this study affected our subjects' latency to respond (at least in the first session), their threat potential may simply not have been strong enough to also impair performance accuracy.

We did not observe a significant relationship between interference effects in Session 1 and personality measures, as they are frequently reported in studies with human participants. While many different explanations are conceivable to explain the absence of an effect, it has to be acknowledged that a sample size of only six subjects implies low statistical power to detect moderator effects of personality variables, if they exist. For future studies that examine to what extent personality moderates the relationship between emotion and cognition in nonhuman primates, larger sample sizes would certainly be desirable. In order to allow for cross-study comparisons, we made our results with regard to personality variables available in spite of this methodological caveat.

General Discussion

Overall, in Session 1 of our modified emotional Stroop task, chimpanzee subjects who had had experience with an anaesthetization procedure responded more slowly in trials presenting them with stimuli depicting the veterinarian than in trials presenting them with other stimuli, and this slow-down effect was more pronounced for subjects whose anaesthetization experience was more recent. As this suggests that stimuli of negative valence impaired performance in a color discrimination task, this effect from Experiment 2 is comparable to the emotional Stroop effect frequently reported in human participants (e.g., Pratto & John, 1991).

Based on our results alone, it is unclear to what extent the chimpanzees recognized the humans (including the veterinarian) depicted in the photographs. It could be argued, for example, that surface perceptual features unique to the veterinarian stimuli (such as color, contrast, contour, etc.) made them more threatening or interesting to look at, and that this, rather than their emotional valence, slowed down responses in veterinarian trials. This would not, however, explain the fact that interference effects were more pronounced for subjects whose last anaesthetization experience was more recent. Secondly, it could be argued that even if it is to be assumed that the chimpanzees did recognize some details in the veterinarian stimuli which then triggered negative associations as a result of the chimpanzees' past experiences, based on our results alone it remains unclear whether it was the identity of the veterinarian or other details such as the blowpipe or the veterinarian's work gear that bore the strongest negative associations and thus were mostly responsible for the slowing of responses. Finally, we acknowledge that our study is not informative with regard to whether such details were recognized as representations of their real life counterparts, or whether they were confused with them (see Fagot, Martin-Malivel, & Depy, 1999). While our study was not designed to investigate these different possibilities, they have no bearing on the main findings of Experiment 2 that the veterinarian

724 stimuli slowed down responding more than any other category of humans, and that the extent of
725 this slowdown varied systematically with the time passed since the last anaesthetization
726 experience.

727 While our results show that presenting our chimpanzee subjects with pictures of the
728 veterinarian slowed down their responding, it remains an open question which stages of
729 executing the task are primarily disrupted by the presence of these stimuli. Identifying which one
730 of the two stimuli is the target may be interrupted, e.g. if the veterinarian stimuli bind attentional
731 resources more strongly than other stimulus categories. It is also conceivable that action
732 execution (touching the selected stimulus) is affected by the presence of veterinarian stimuli, as
733 negative stimuli are usually avoided rather than approached. In this case, the slowing of
734 responses would reflect a reluctance to touch an aversive stimulus that has already been
735 identified as the target, rather than a binding of attentional resources that disrupts target
736 identification. This possibility could be ruled out in future studies if the stimuli are not present
737 during the time of action execution (e.g. by presenting the picture stimuli embedded in the color
738 frames only for a brief period before either stimulus is touched). Ambiguity with regard to which
739 cognitive processes are involved in the slowdown has also been the subject of extensive debates
740 over the interpretation of emotional Stroop effects in the human literature (e.g. Algom et al.,
741 2004; de Ruiter & Brosschot, 1994; MacLeod et al., 1986; Yiend et al., 2013).

742 In conclusion, we propose our modified version of the emotional Stroop task as an easily
743 implemented method to study the relationship between emotion and cognition in nonhuman
744 primates, and possibly other species. However, considering the limitations with regard to the
745 interpretability of interference effects, we agree with Yiend et al. (2013) that the paradigm may
746 not be the best method to study *how* emotional stimuli disrupt cognitive task execution. If the

strong effect that emotional valence had on task performance in this study can be replicated in future studies, we recommend the task instead as a method to study *which* stimuli (or stimulus categories) interfere with cognitive performance by virtue of their emotional valence. It may also offer a possibility to study how individuals differ with regard to how much their performance is affected. In this sense, the task may be suitable as a diagnostic tool to measure anxiety with regard to particular stimuli at the group or individual level, e.g. to investigate relationships between individuals from the same group.

References

Algom D, Chajut E, Lev S (2004) A rational look at the emotional stroop phenomenon: a generic slowdown, not a stroop effect. J Exp Psychol Gen 133:323-338. doi:10.1037/0096-3445.133.3.323

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. Psychol Bull 133:1-24. doi:10.1037/0033-2909.133.1.1

Bateson M, Nettle D (2015) Development of a cognitive bias methodology for measuring low mood in chimpanzees. PeerJ 3:e998. doi: 10.7717/peerj.998

Ben-Haim MS, Mama Y, Icht M, Algom D (2014) Is the emotional Stroop task a special case of mood induction? Evidence from sustained effects of attention under emotion. Atten Percept Psychophys 76:81-97. doi:10.3758/s13414-013-0545-7

Bethell, EJ (2015). A "how-to" guide for designing judgment bias studies to assess captive animal welfare. J Appl Anim Welf Sci, 18(sup1): 18-42. doi: 10.1080/10888705.2015.1075833

Bethell EJ, Holmes A, Maclarnon A, Semple S (2012) Cognitive bias in a non-human primate: husbandry procedures influence cognitive indicators of psychological well-being in captive rhesus macaques. Anim Welf 21:185–195. doi:10.7120/09627286.21.2.185

Bovet D, Vauclair J (2000) Picture recognition in animals and humans. Behav Brain Res 109:143-165. doi:10.1016/S0166-4328(00)00146-7

Boysen ST, Berntson GG (1989) Conspecific recognition in the chimpanzee (Pan troglodytes): cardiac responses to significant others. J Comp Psychol 103:215-220. doi:10.1037/0735-7036.103.3.215

778 Buckley TC, Blanchard EB, Neill WT (2000) Information processing and PTSD: A review of the
779 empirical literature. Clin Psychol Rev 20:1041-1065. doi:10.1016/S0272-7358(99)00030-
780 6
781 Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral
782 sciences. Hillsdale, NJ: Erlbaum
783 Constantine R, McNally RJ, Hornig CD (2001) Snake fear and the pictorial emotional Stroop
784 paradigm. Cognitive Ther Res 25:757-764. doi:10.1023/A:1012923507617
785 De Ruiter C, Brosschot JF (1994) The emotional Stroop interference effect in anxiety: attentional
786 bias or cognitive avoidance? Behav Res Ther 32:315-319. doi:10.1016/0005-
787 7967(94)90128-7
788 Fagot J, Martin-Malivel J, Dépy D (2000) What is the evidence for an equivalence between
789 objects and pictures in birds and nonhuman primates. In: Fagot J (ed) Picture perception
790 in animals. Psychology Press, New York, NY, US, pp 295-320
791 Field M, Cox WM (2008) Attentional bias in addictive behaviors: a review of its development,
792 causes, and consequences. Drug Alcohol Depen 97:1-20.
793 doi:10.1016/j.drugalcdep.2008.03.030
794 Fox E, Griggs L, Mouchlianitis E (2007) The detection of fear-relevant stimuli: Are guns noticed
795 as quickly as snakes? Emotion 7:691-696. doi:10.1037/1528-3542.7.4.691
796 Frings C, Englert J, Wentura D, Bermeitinger C (2010) Decomposing the emotional Stroop
797 effect. Q J Exp Psychol 63:42-49. doi:10.1080/17470210903156594
798 Hirata S et al. (2013) Brain response to affective pictures in the chimpanzee. Sci Rep 3
799 doi:10.1038/srep01342
800 Kano F, Tanaka M, Tomonaga M (2008) Enhanced recognition of emotional stimuli in the

chimpanzee (Pan troglodytes). Anim Cogn 11:517-524. doi:10.1007/s10071-008-0142-7

Kindt M, Brosschot JF (1997) Phobia-related cognitive bias for pictorial and linguistic stimuli. J Abnorm Psychol 106:644-648. doi:10.1037/0021-843X.106.4.644

King HM, Kurdziel LB, Meyer JS, Lacreuse A (2012) Effects of testosterone on attention and memory for emotional stimuli in male rhesus monkeys. Psychoneuroendocrino 37:396-409. doi:10.1016/j.psyneuen.2011.07.010

King JE, Figueredo AJ (1997) The five-factor model plus dominance in chimpanzee personality. J Res Pers 31:257-271. doi:10.1006/jrpe.1997.2179

Koda H, Sato A, Kato A (2013) Is attentional prioritisation of infant faces unique in humans?: Comparative demonstrations by modified dot-probe task in monkeys. Behav Process 98:31-36. doi:10.1016/j.beproc.2013.04.013

Koster EH, Crombez G, Van Damme S, Verschuere B, De Houwer J (2004) Does imminent threat capture and hold attention? Emotion 4:312-317. doi:10.1037/1528-3542.4.3.312

Lacreuse A, Schatz K, Strazzullo S, King HM, Ready R (2013) Attentional biases and memory for emotional stimuli in men and male rhesus monkeys. Anim Cogn 16:861-871. doi:10.1007/s10071-013-0618-y

Lang PJ, Davis M, Öhman A (2000) Fear and anxiety: animal models and human cognitive psychophysiology. J Affect Disord 61:137-159. doi:10.1016/S0165-0327(00)00343-8

Lavy E, Van den Hout M (1993) Selective attention evidenced by pictorial and linguistic Stroop tasks. Behav Ther 24:645-657. doi:10.1016/S0005-7894(05)80323-5

Mackay DG, Shafto M, Taylor JK, Marian DE, Abrams L, Dyer JR (2004) Relations between emotion, memory, and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. Mem Cognition 32:474-488. doi:10.3758/BF03195840

824 MacLeod C, Mathews A, Tata P (1986) Attentional bias in emotional disorders. J Abnorm
825 Psychol 95:15-20. doi:10.1037//0021-843X.95.1.15
826 Marzouki Y, Gullstrand J, Goujon A, Fagot J (2014) Baboons' response speed is biased by their
827 mood. PLOS ONE 9(7): e102562. doi:10.1371/journal.pone.0102562
828 Mathews A, MacLeod C (1985) Selective processing of threat cues in anxiety states. Behav Res
829 Ther 23:563-569. doi:10.1016/0005-7967(85)90104-4
830 Mauer N, Borkenau P (2007) Temperament and early information processing: Temperament-
831 related attentional bias in emotional Stroop tasks. Pers Indiv Differ 43:1063-1073.
832 doi:10.1016/j.paid.2007.02.025
833 McKenna FP, Sharma D (2004) Reversing the emotional Stroop effect reveals that it is not what
834 it seems: the role of fast and slow components. J Exp Psychol Learn 30:382-392.
835 doi:10.1037/0278-7393.30.2.382
836 Mogg K, Bradley BP (2003) Selective Processing of Nonverbal Information in Anxiety:
837 Attentional Biases for Threat. In: Philippot P, Feldman RS, Coats EJ (eds) Nonverbal
838 behavior in clinical settings. Oxford University Press, New York, NY, US, pp 127-143
839 Öhman A, Mineka S (2001) Fears, phobias, and preparedness: toward an evolved module of fear
840 and fear learning. Psychol Rev 108:483-522. doi:10.1037/0033-295X.108.3.483
841 Parr LA (2001) Cognitive and physiological markers of emotional awareness in chimpanzees
842 (Pan troglodytes). Anim Cogn 4:223-229. doi:10.1007/s100710100085
843 Paul ES, Harding EJ, Mendl M (2005) Measuring emotional processes in animals: the utility of a
844 cognitive approach. Neurosci Biobehav Rev 29:469-491.
845 doi:10.1016/j.neubiorev.2005.01.002
846 Phaf RH, Kan K-J (2007) The automaticity of emotional Stroop: A meta-analysis. J Behav Ther

847 Exp Psy 38:184-199. doi:10.1016/j.jbtep.2006.10.008
848 Pomerantz O, Terkel J, Suomi SJ, Paukner A (2012) Stereotypic head twirls, but not pacing, are
849 related to a 'pessimistic'-like judgment bias among captive tufted capuchins (Cebus
850 apella). Anim Cogn 15:689–698. doi: 10.1007/s10071-012-0497-7
851 Pratto F, John OP (1991) Automatic Vigilance: The Attention-Grabbing Power of Negative
852 Social Information. J Pers Soc Psychol 61:380-391. doi:10.1037/0022-3514.61.3.380
853 Robbins SJ, Ehrman RN (2004) The role of attentional bias in substance abuse. Behav Cogn
854 Neurosci Rev 3:243-260. doi:10.1177/1534582305275423
855 Schmidt LJ, Belopolsky AV, Theeuwes J (2014) Attentional capture by signals of threat.
856 Cognition Emotion doi:10.1080/02699931.2014.924484
857 Shibasaki M, Isomura T, Masataka N (2014) Viewing images of snakes accelerates making
858 judgements of their colour in humans: red snake effect as an instance of 'emotional
859 Stroop facilitation'. R Soc Open Sci 1 doi:10.1098/rsos.140066
860 Shibasaki M, Kawai N (2009) Rapid detection of snakes by Japanese monkeys (Macaca fuscata):
861 an evolutionarily predisposed visual system. J Comp Psychol 123:131-135.
862 doi:10.1037/a0015095
863 Siegrist M (1995) Effects of taboo words on color-naming performance on a Stroop test. Percept
864 Motor Skill 81:1119-1122. doi:10.2466/pms.1995.81.3f.1119
865 Underwood AJ (1997) Experiments in ecology: their logical design and interpretation using
866 analysis of variance. Cambridge University Press, Cambridge
867 Waters AJ, Sayette MA, Wertz JM (2003) Carry-over effects can modulate emotional Stroop
868 effects. Cognition Emotion 17: 501-509. doi:10.1080/02699930143000716
869 Weiss A et al. (2009) Assessing chimpanzee personality and subjective well-being in Japan. Am

J Primatol 71:283-292. doi:10.1002/ajp.20649

Williams JMG, Mathews A, MacLeod C (1996) The emotional Stroop task and psychopathology. Psychol Bull 120:3-24. doi:10.1037/0033-2909.120.1.3

Witthöft M, Rist F, Bailer J (2008) Enhanced early emotional intrusion effects and proportional habituation of threat response for symptom and illness words in college students with elevated health anxiety. Cognitive Ther Res 32:818-842. doi:0.1007/s10608-007-9159-5

Yiend J (2010) The effects of emotion on attention: A review of attentional processing of emotional information. Cognition Emotion 24:3-47. doi:10.1080/02699930903205698

Yiend J, Barnicot K, Koster EH (2013) Attention and emotion. In: Robinson MD, Watkins ER, Harmon-Jones E (eds) Handbook of cognition and emotion. Guilford Press, New York, NY, USA, pp 97-116

Table 1.

Subject	Sex	Age	Categorization Training A	Categorization Training B	Additional Categorization Training A	Transfer Test
Kofi	male	7	7	-	-	2
Riet	female	35	14	-	-	2
Lobo	male	9	17	-	-	2
Lome	male	11	19	-	-	2
Tai	female	10	28	-	-	2
Fraukje	female	37	40	-	-	2
Sandra	female	19	(40)	5	3	3
Kara	female	7	(40)	13	8	6
Robert	male	37	(40)	(40)	-	-
Corrie	female	36	(40)	(40)	-	-

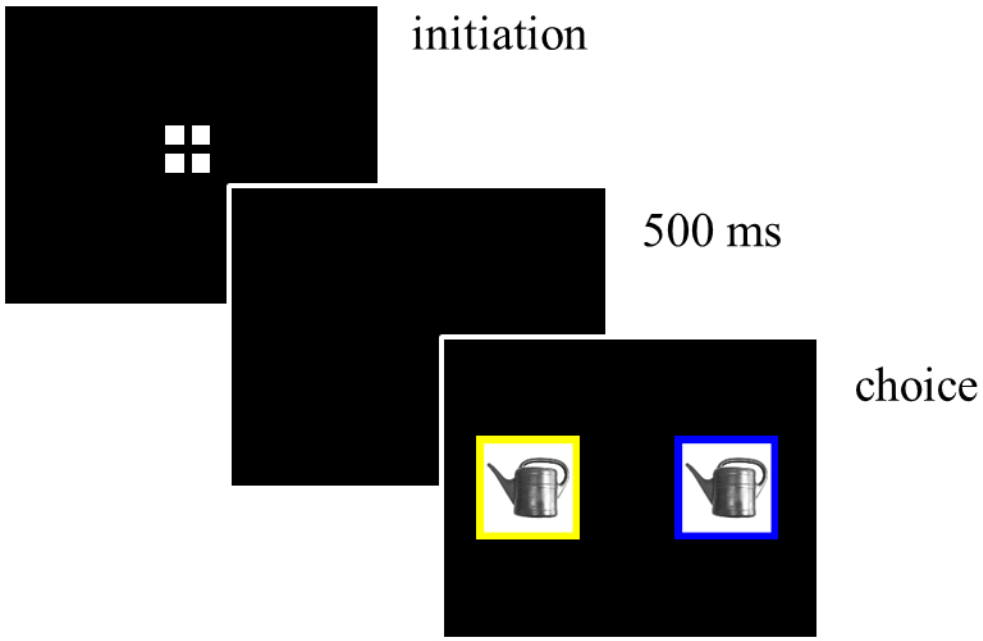
Numbers in parentheses indicate that criterion was not reached within reported number of sessions.

881

Table 2.

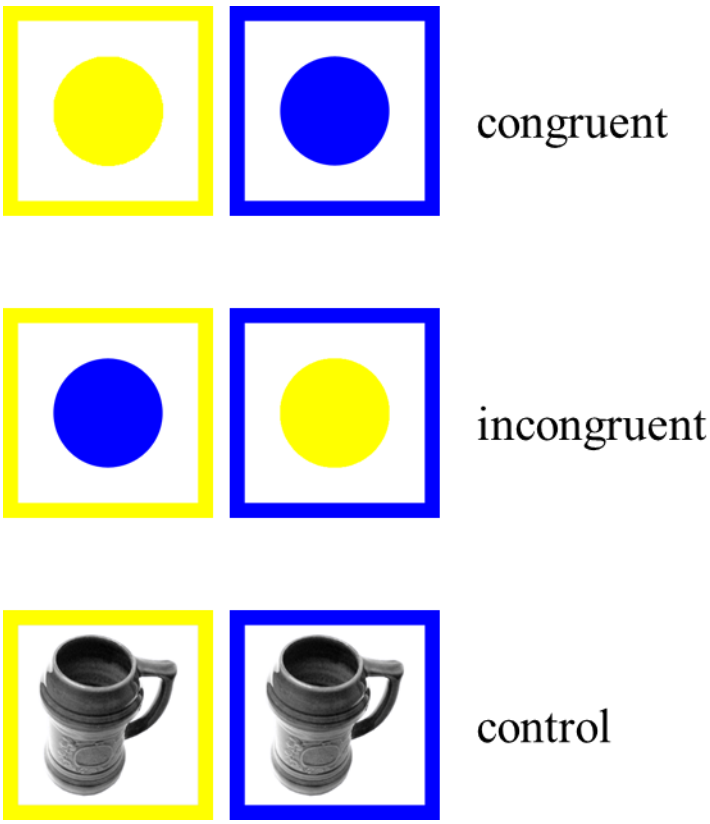
Category	Description
control	white square
veterinarian	veterinarian (face)
veterinarian	veterinarian (face, wearing face mask and hairnet cap)
veterinarian	veterinarian (from the waist up, holding blow pipe)
veterinarian	veterinarian (from the waist up, holding blow pipe, wearing face mask and hair net cap)
caretaker	caretaker 1 (face)
caretaker	caretaker 2 (face)
caretaker	caretaker 1 (from the waist up, wearing zoo work gear, holding food bucket)
caretaker	caretaker 2 (from the waist up, wearing zoo work gear, holding food bucket)
stranger	stranger 1 (face)
stranger	stranger 2 (face)
stranger	stranger 1 (from the waist up, holding backpack)
stranger	stranger 2 (from the waist up, holding backpack)

883 Fig 1a.



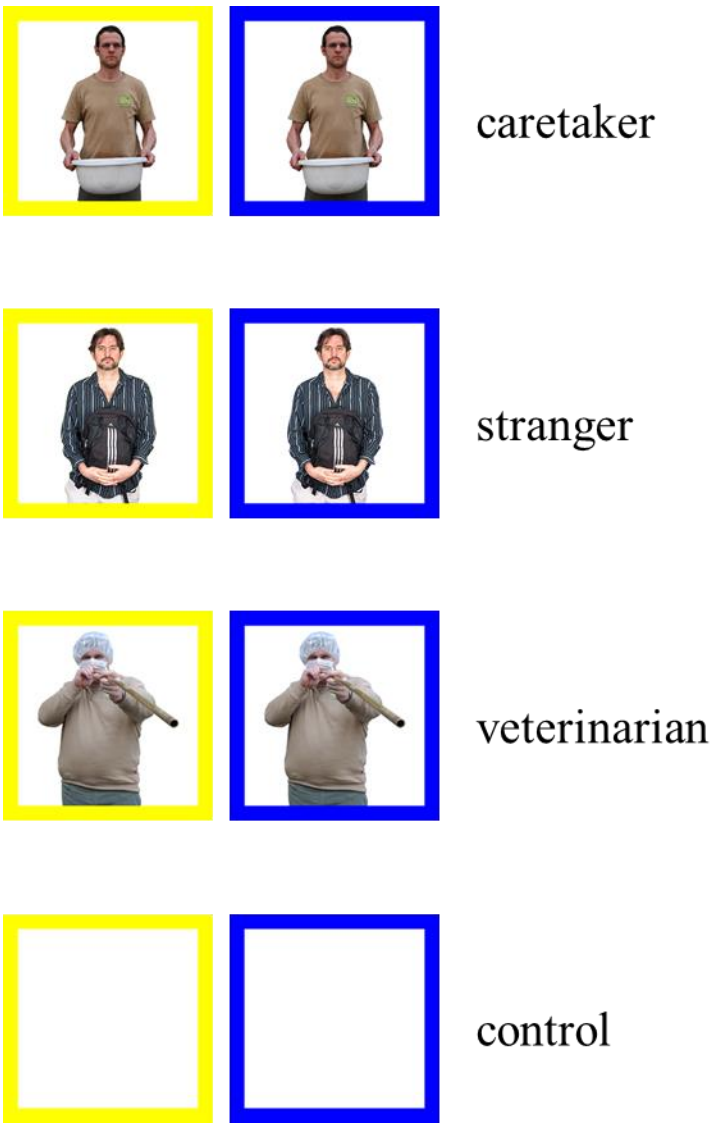
884

885 Fig 1b.



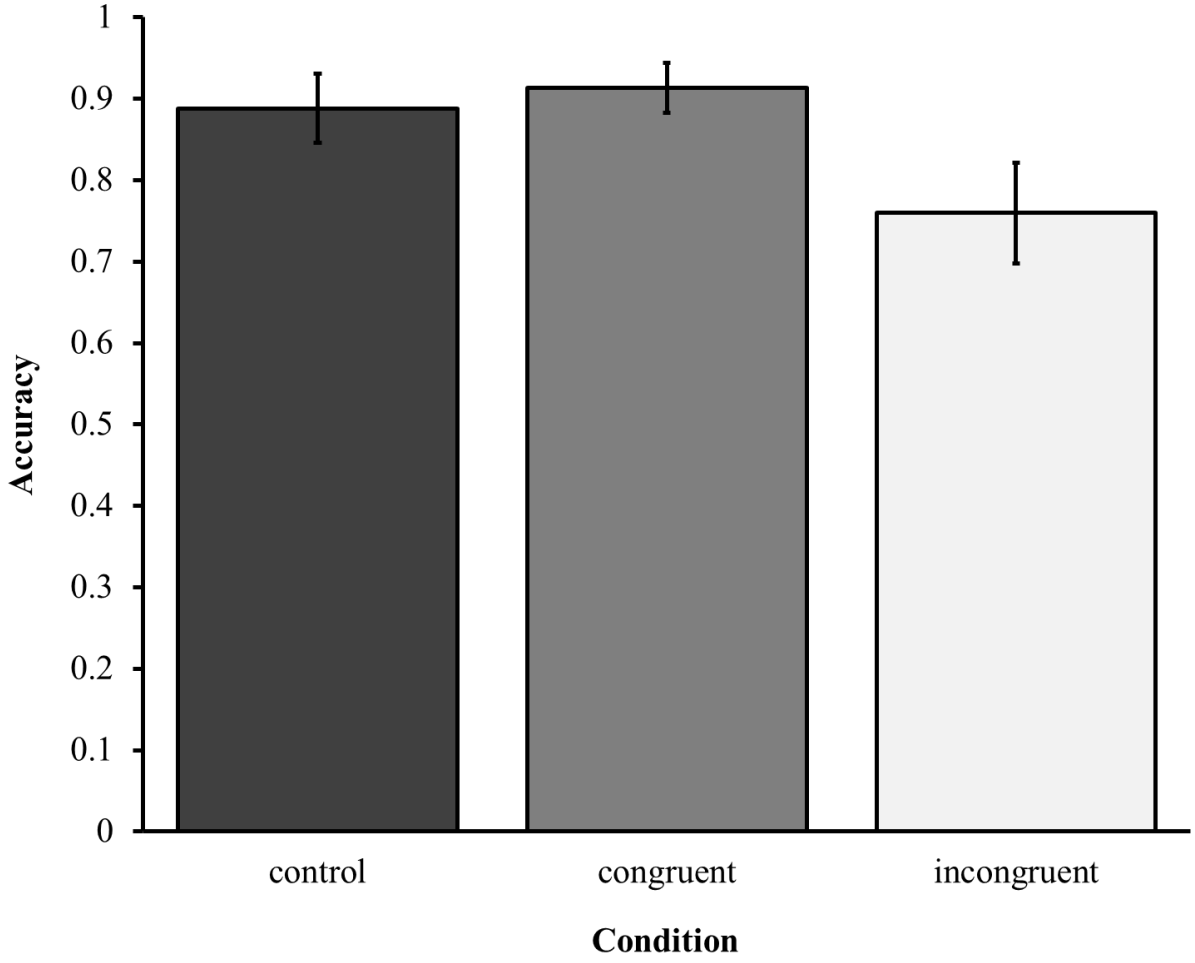
886

887 Fig 1c.



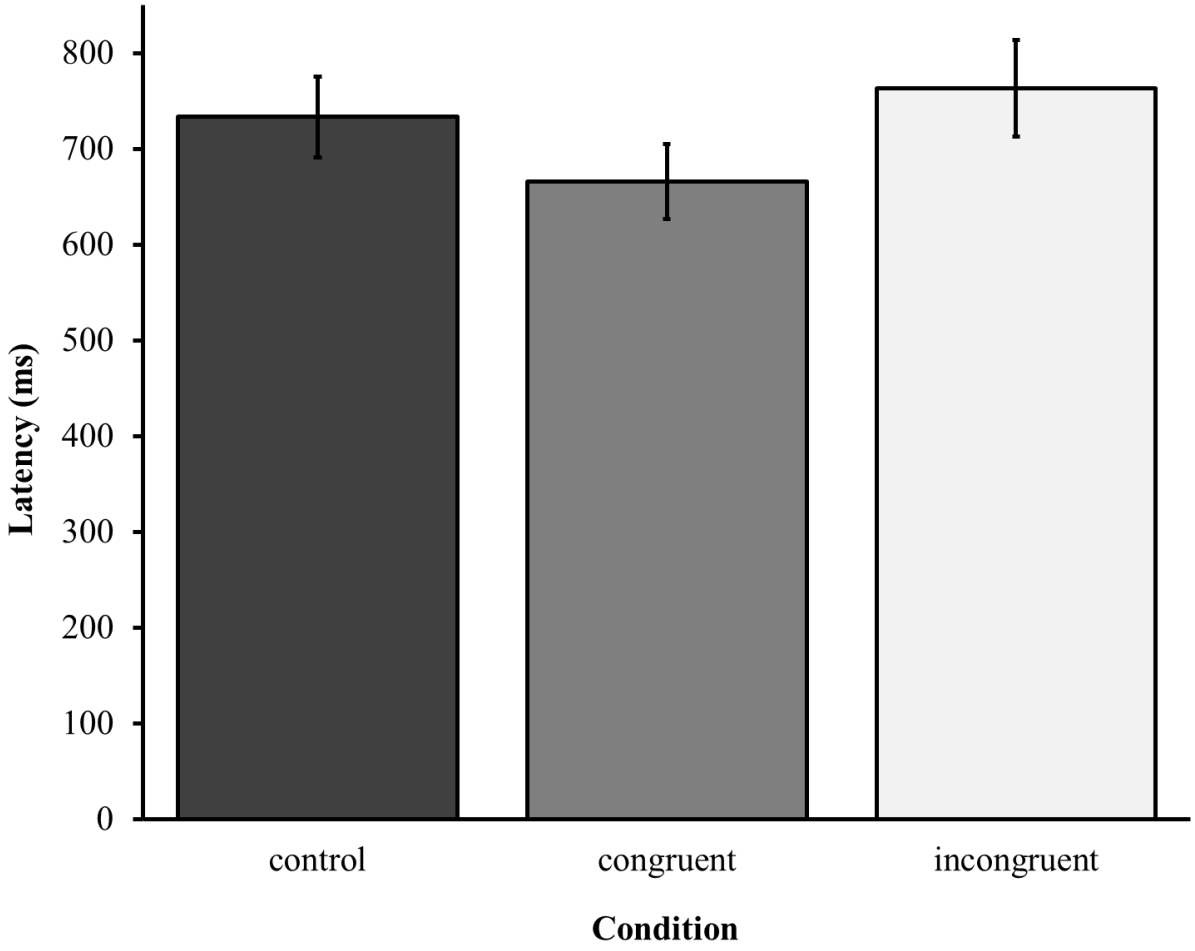
888

889 Fig 2a.



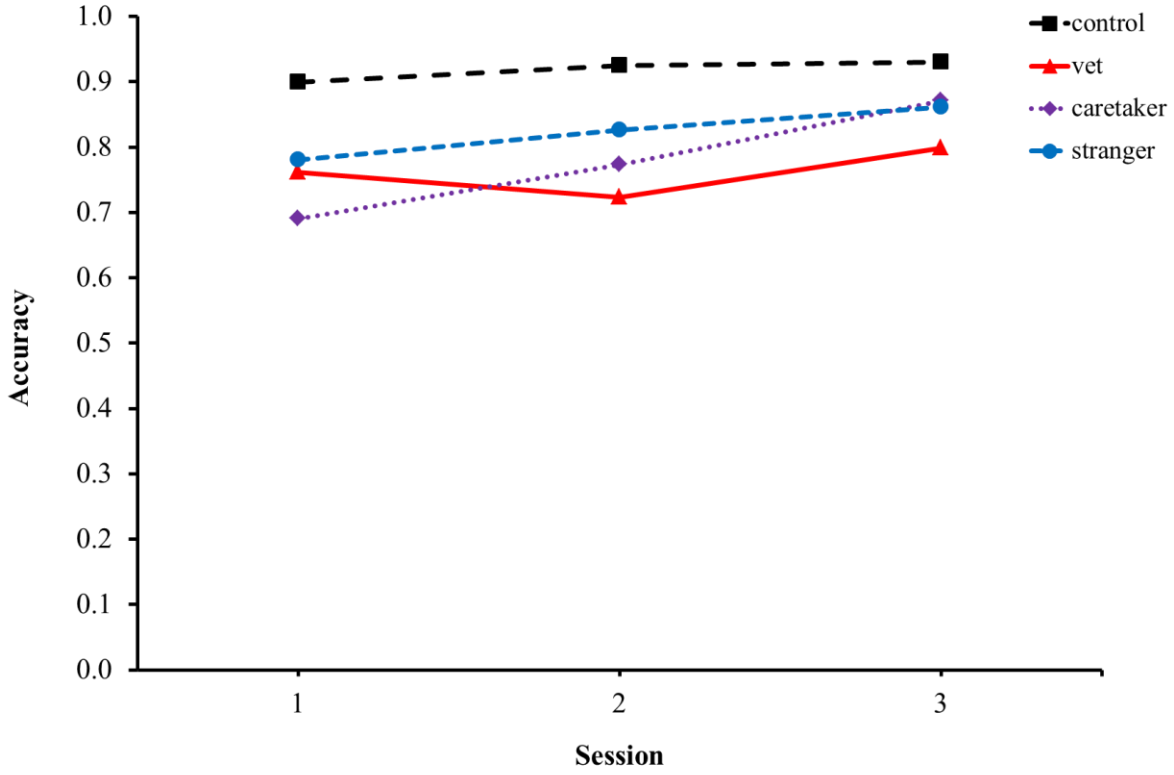
890

891 Fig 2b.



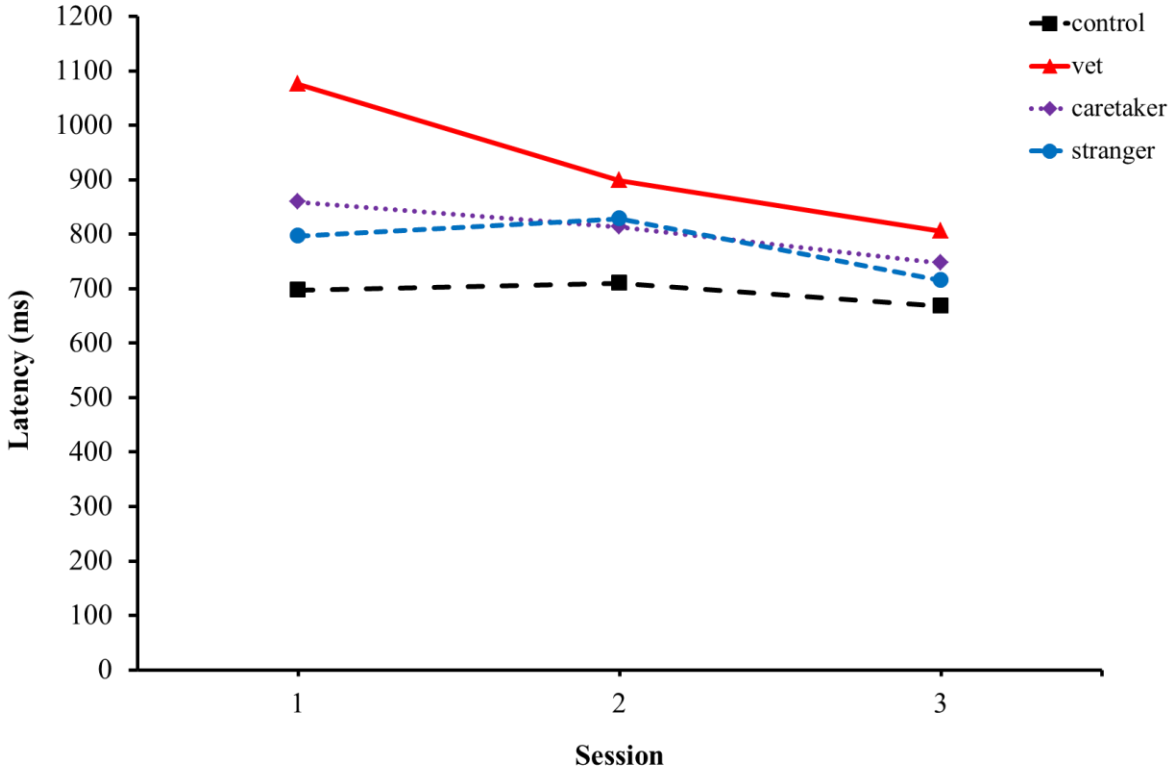
892

893 Fig 3a.



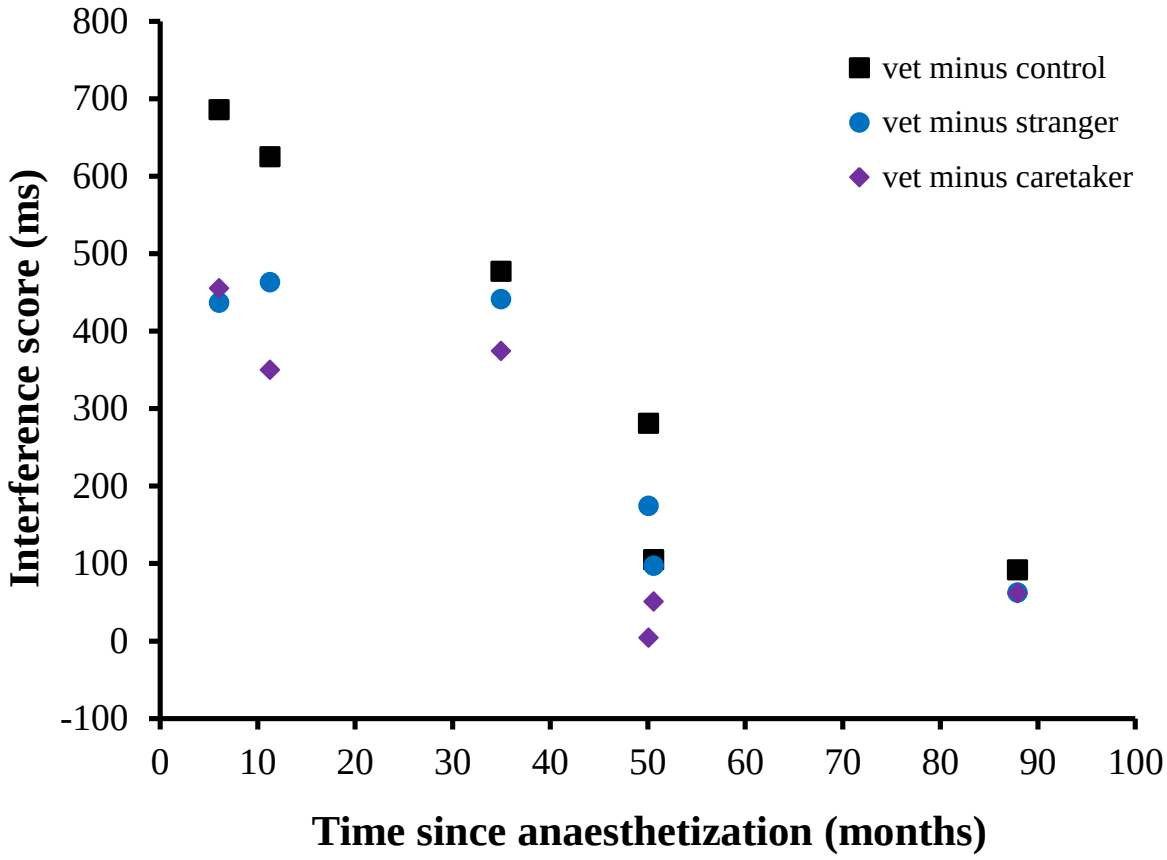
894

895 Fig 3b.



896

897 Fig 3c.



898

Table captions and figure legends

Table 1 Number of sessions required to reach criterion in each training condition

Table 2 Stimuli used in Experiment 2

Figure 1a Trial procedure for color discrimination training A, transfer test, and Experiments 1 and 2. The figure depicts a control trial from Experiment 1.

Figure 1b Example stimuli from Experiment 1.

Figure 1c Example stimuli from Experiment 2.

Figure 2a Accuracy in different conditions in Experiment 1. Error bars represent SEM.

Figure 2b Latency in different conditions in Experiment 1. Error bars represent SEM.

Figure 3a Accuracy in different conditions across sessions in Experiment 2.

Figure 3b Latency in different conditions across sessions in Experiment 2.

Figure 3c Interference scores (response latency differences between vet condition and other conditions in Session 1 of Experiment 2) as a function of time since the last anaesthetization

Footnotes

1 It should be noted that for the color discrimination training A, the onsets of the trial initiation display, the 500ms waiting display, the stimuli, and the feedback interval were each accompanied by additional program-execution-related average delays of approximately 1 to 16 ms. For the color discrimination training B, the onsets of the trial initiation display and of the 500ms waiting display, the first onset of the stimuli, and the feedback interval onset were each accompanied by additional program execution related average delays of approximately 7 to 16 ms.

2 Following the suggestion of an anonymous reviewer, we also explored latencies in incorrect trials. Across sessions and categories, while response latencies in our sample of seven subjects tended to be slower in incorrect trials (mean of latency medians: M = 862.00 ms) than in correct trials (M = 768.07 ms), this effect was not statistically significant, $t(6) = 1.55$, $p = .173$. We further investigated specifically for trials presenting the veterinarian, whether response latencies from subjects with anaesthetization experience were slower in incorrect than in correct trials: considering data from all three sessions, we did not find a significant difference between latencies in correct vs. incorrect vet trials, $t(5) = 1.43$, $p = .211$. Considering Session 1 alone, in spite of a sizable mean difference in response latency between incorrect vet trials (M = 1320.08 ms) and correct vet trials (M = 908.75 ms), this effect was not statistically significant, $t(5) = 1.98$, $p = .105$. We would like to add a note of caution with regard to these results, however. For several subjects the rate of incorrect responses was very low. In particular, when considering vet trials alone, this means that some of the latency scores that had to be used in these analyses were based on as little as two to four data points (three subjects in Session 1). Such small numbers imply low measurement reliability, even when medians are used as measures of central tendency.

For the same reason, latency comparisons between different categories that were restricted to incorrect trials were not carried out because several subjects made no errors in one or more of the non-veterinarian categories in one or more of the three sessions.