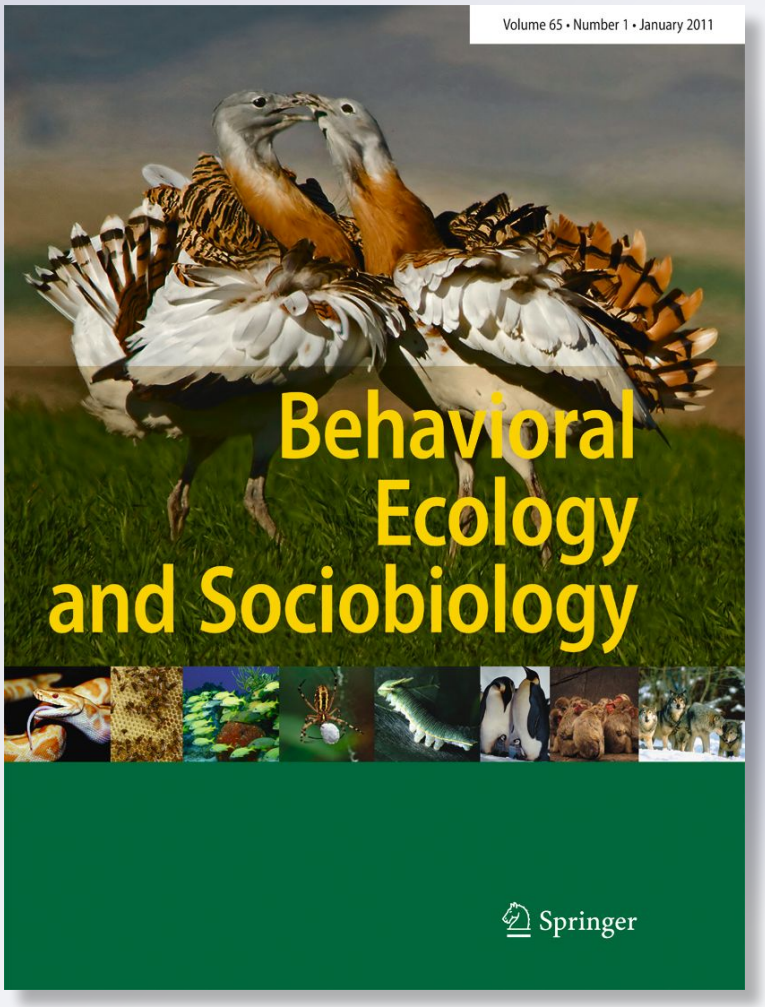
Evaluating animal personalities: do observer assessments and experimental tests measure the same thing?

Alecia J. Carter, Harry H. Marshall, Robert Heinsohn & Guy Cowlishaw

Behavioral Ecology and Sociobiology

ISSN 0340-5443
Volume 66
Number 1

Behav Ecol Sociobiol (2012) 66:153-160
DOI 10.1007/s00265-011-1263-6



Springer

Behav Ecol Sociobiol (2012) 66:153–160
DOI 10.1007/s00265-011-1263-6

ORIGINAL PAPER

Evaluating animal personalities: do observer assessments and experimental tests measure the same thing?

Alecia J. Carter · Harry H. Marshall · Robert Heinsohn · Guy Cowlishaw

Received: 26 July 2011 / Revised: 8 September 2011 / Accepted: 10 September 2011 / Published online: 29 September 2011

Abstract The animal personality literature uses three approaches to assess personality. However, two of these methods, personality ratings and experimentation, have been little compared in captivity and never compared in the wild. We assessed the boldness of wild chacma baboons *Papio ursinus* using both ratings and experimental methods. Boldness was experimentally assessed when individuals were presented with a novel food item during natural foraging. The boldness of the same individuals was rated on a five-point scale by experienced observers. The ratings and experimental assessments of boldness were found to correlate positively and in a linear fashion. When considered categorically the two approaches showed variable agreement depending on the number of categories assigned and the cut-off criteria adopted. We suggest that the variation between approaches arises because each method captures different aspects of personality; ratings consider personality in absolute terms (using predefined criteria) and multiple contexts, while experimental assessments consider personality in relative terms (using experimental scores relative to the population average) and in limited contexts. We encourage animal personality researchers to consider adopting both methodologies in future studies. We also propose that future studies restrict their analyses to continuous data, since the greatest comparability between methods was found with these data. However, if individuals must be categorised, we suggest that researchers either (a) analyse only those individuals categorised as bold or shy by both ratings and experimental approaches or, if these methods cannot be employed simultaneously, (b) do not use approach-specific criteria but choose a cut-off that can be compared by both approaches.

Keywords Behavioural syndromes · Boldness · Chacma baboon · Personality

Communicated by T. Bakker

A. J. Carter (✉) · R. Heinsohn
The Fenner School of Environment and Society,
The Australian National University,
Acton,
Canberra, ACT 0200, Australia
e-mail: alecia.carter@anu.edu.au

A. J. Carter · H. H. Marshall · G. Cowlishaw
The Institute of Zoology, Zoological Society of London,
Regent's Park,
London NW1 4RY, UK

H. H. Marshall
Division of Biology, Imperial College London,
Silwood Park,
Berkshire SL5 7PY, UK

Introduction

Personality in animals refers to consistent and repeatable behaviour at the level of the individual (Sih et al. 2004). The evolutionary origins and ecological significance of personality variation in animals is a rapidly expanding area of research. However, speed of progress in this field may be hampered by the widespread use of three different methodological approaches: subjective personality ratings, behavioural coding and experimentation. Subjective assessments use ratings of multiple items, such as adjectives or behavioural descriptors, by observers familiar with individual animals to describe the dimensions encompassing multiple personality axes. Behavioural codings consist of recording the behaviour of a focal individual according to a predetermined ethogram, that is, a list of discrete behaviours performed by the species under study during the animal's natural behaviour (Gosling 2001; Vazire et al.

2007). Experimental assessments similarly record individual patterns of behaviour, but in response to controlled experimental stimuli, to assess variability in a 'limited' number of personality axes such as boldness, aggressiveness or sociability (see Sih et al. 2004; for a review, Nettle and Penke 2010). All three methodological approaches tend to use data reduction methods, such as principle components or factor analysis, to translate their data into one or more personality traits for the study species. Combinations of these approaches are frequently used by personality researchers in the fields of comparative psychology and behavioural ecology, to varying degrees. Behavioural ecologists frequently use behavioural codings of natural behaviour and behaviour during experimentation to assess personality. Comparative psychologists often use ratings and behavioural codings of natural behaviour (hereafter, natural behaviour) but also employ experimentation to assess personality. In all methodological approaches, the derived personality scores reflect an individual's position on the relevant behavioural continuum, such as a shy–bold axis, but can also be used to assign an individual to a discrete group, e.g. shy, bold or intermediate (for example, van Oers et al. 2005; Kurvers et al. 2010; Sinn et al. 2008).

Although the relationships between the different approaches are well established between natural behaviour and experimentation, and between natural behaviour and ratings (for examples, see Pederson et al. 2005; Konecna et al. 2008; Carter et al. 2010; Kurvers et al. 2010), there has been little comparison between ratings and experimentation (Itoh 2002; Freeman and Gosling 2010), resulting in a lack of convergent validity for these measures (Campbell and Fiske 1959). This is surprising, since an understanding of the correspondences and differences between these two methods would be of great value to those who use these approaches. The only exceptions we know of are from captive studies and generally consist of relatively small sample sizes (for examples, see Table 1). The authors of these studies directly compared ratings and experimental scores (using a multi-trait, multimethod approach; see Bergvall et al. 2011 for an example), and while they found positive relationships in many of the personality traits investigated, they were unable to do so in others. Indeed, Uher and Asendorpf (2008) suggested that adjective ratings may not be the most accurate at predicting manifest behaviour in their study, thus highlighting the need for more comparative work. In addition, these studies did not investigate whether this relationship applies in wild animals (captivity can affect the expression and evolution of animal personality, McDougall et al. 2006). Finally, no studies have yet explored the consistency of categorical personality assignments between the two approaches.

In this study we use a single-trait, multimethod approach to compare assessments of boldness using ratings and experiments in a wild population (Campbell and Fiske 1959). We chose to focus on boldness as it is the most studied personality axis in the behavioural ecology literature (Sih et al. 2004). Our analysis takes two steps. First, we examine whether observer ratings and experimental scores on the bold–shy continuum are correlated. Second, to further investigate this relationship, we explore whether the categorisation of individuals as bold or shy is consistent between the ratings and experimental methods using different threshold criteria. The latter will also allow us to improve the comparability of results in future studies by making recommendations regarding the most robust cut-off criteria for categorisation.

Materials and methods

Study area and species

We studied chacma baboons *Papio ursinus* in 2009 at the Tsaobis Leopard Park, Namibia (15°45′ E, 22°23′ S). Two groups of chacma baboons (n=44, 31) have been habituated to the presence of observers at close range and are individually recognisable (see Huchard et al. 2010 for general methods of behavioural data collection at this site). All adult, subadult and juvenile baboons were tested (we did not test dependent infants under 1 year of age), resulting in a total of 58 individuals for the following analyses.

Boldness ratings

Seven observers rated the boldness of each individual baboon on a five-point scale. On this scale, −2 indicated that the baboon was *shy*, retreats readily from others or outside disturbances and can be over-vigilant or fearful; and 2 indicated that the baboon was *bold*, behaves in a positive, assured and bold manner, not restrained or tentative. All observers had been following the baboon troops for at least 3 months (maximum 6 months) between May and November 2009, although one observer followed only one group and thus rated only individuals from that group. Observers spent a minimum of 12 h a day, on average 3 days in every five, following the baboons and collecting detailed behavioural data. They were thus familiar with most individuals. However, due to different protocols followed by different observers (for other research purposes beyond the present study), two observers were not confident assessing some baboons and thus did not score all individuals. Each baboon was rated by at least four observers (mean=5.9, median, mode=6, range=4–7 observers). Observers did not discuss their ratings with each other.

Table 1 Research directly comparing the relationship between experiments and observer ratings of animal personality

Study	Number	Study species	Significant correlations	Test used[a]
Carlstead et al. (1999)	53	Black rhinoceros (*Diceros bicornis*)	0.25–0.62[b, c]	*S* cor
Wielebnowski (1999)	44	Cheetah (*Acinonyx jubatus*)	0.26–0.39[b, d]	*S* cor
Uher and Asendorpf (2008)	20[e]	Great apes	0.56	*P* cor
Bergvall et al. (2011)	15	Fallow deer (*Dama dama*)	0.79[f]	*S* cor
Carter et al. (this study)	58	Chacma baboon (*P. ursinus*)	0.56[f]	*P* cor

[a] *P* cor indicates studies that used a Pearson correlation, *S* cor indicates studies that used Spearman rank correlations

[b] No data reduction methods were used as in the other studies; numbers represent the range of significant ($P<0.05$) correlations between various experimental responses and personality items found through observer ratings

[c] This study compared six personality items with eight experimental responses to obtain 12 significant correlations from 48 (25%) possible correlations

[d] This study compared 15 personality items with five experimental responses to obtain 22 significant correlations from 75 (29%) possible correlations

[e] The sample per taxon comprised five each of bonobos (*Pan paniscus*), chimpanzees (*Pan troglodytes verus*), gorillas (*Gorilla gorilla gorilla*) and orangutans (*Pongo abelii*)

[f] Correlation between observer ratings and experimental boldness

Boldness experiment

Each baboon was presented with a novel food item (akin to novel object tests; for example, see Bremner-Harrison et al. 2004) while foraging under natural conditions. Baboons were presented with the novel food when they were travelling between food patches and had no neighbours within 25 m. If an individual baboon picked up the novel food and moved with it into the presence of other baboons, those individuals were noted and did not receive that food type when tested. Novel food types comprised hard-boiled eggs with the shell on or removed then coloured with food dye (red or green, Moir's food dye), or small egg-shaped bread rolls either non-dyed or coloured as with the eggs. All experiments were filmed to facilitate data extraction (Panasonic SDR-SW20, Kadoma Osaka, Japan; see online resource 1). All experiments were performed before 1,200 h. Observers that completed the boldness ratings did not witness any of the boldness experiments.

The following data were extracted from the videos: latency to approach food item on detection (in seconds; if the food item was not approached, the individual was given the maximum value of 150 s; individuals that did not detect the item were subsequently retested), time spent inspecting the food item (in seconds; the time between approaching the food item and the end of the experiment, either leaving or eating the item), latency to handle the food item (in seconds; max. 150 s), latency to consume the food item (in seconds; max. 150 s) and time spent handling the food item (in seconds; the time spent touching the food item).

Statistical analyses

We obtained data from all 58 adult, subadult and juvenile male and female baboons present throughout the study. For our personality ratings, we calculated the median of the observer ratings of boldness for each baboon, generating personality scores on a nine-point scale (allowing for midrange values on our five-point assessment scale). Inter-rater agreement was calculated using the intra-class correlation coefficient (ICC 3, *k*; Shrout and Fleiss 1979). For our experimental personality test, the five variables from the novel food experiments were standardized to have a mean of 0 and a standard deviation of 1 and then reduced to a smaller number of variables (dimensions) using principal components analysis using the differentially weighted scores for further analysis. The first two dimensions, principal components PC1 and PC2, were retained for investigation following parallel analysis (Horn 1965). Our data analysis then took two steps.

First, we investigated the relationship between the experimental scores and the observer ratings. The relationships between the median rating and both PC1 and PC2 were analysed using a linear mixed effects model (LMM) in which the former was the response variable and the latter a fixed effect. Since previous studies of primate personality have found a significant effect of age on personality dimensions such as extraversion, agreeableness and opportunism (McGuire et al. 1994; Weiss et al. 2007), age class was also included, as a random effect. We also tested for an effect of troop, sex and food type presented on the median rating, by comparing (by a likelihood ratio test) the fit of the data to a null model (a linear model) with the four linear mixed effects models where troop, sex or food type were included as random effects, respectively. The residuals were normally distributed and not overdispersed.

Second, we explored how consistently individuals were assigned to bold/shy categories between the two approaches. That is, if an individual is categorised as bold or shy by observer rating, what proportion of the time

would experimental assessments also categorise that individual as bold or shy? We did this in two ways: (1) by comparing the consistency between shy/bold categorisations (i.e. providing a clear focus on alternative personality types, but ignoring those individuals consistently classified as intermediate between these two extremes) and (2) by comparing consistency between shy/intermediate/bold categories (thus accounting for the classification of all individuals). The former approach is likely to be of more interest to those working on the behavioural consequences of personality (for example, Carere et al. 2005), while the latter approach will be more useful to those investigating the diversity and distribution of personality types (for example, Bell 2005).

Individuals were assigned a category according to the criteria outlined below. In experimental assessments of personality, individuals are categorised based on their responses in relation to the population response (for example, see van Oers et al. 2005). Thus, we divided our experimental scores (PC1 only) at the median, with 'bold' and 'shy' individuals categorised as higher or lower than the median score, respectively. In contrast, in observer ratings of boldness, individuals are categorised according to their position on the assessment scale. Thus, we assigned individuals scoring higher than the midrange on our ratings scale (in this case 0) as bold and those scoring lower than the midrange as shy (individuals at the midrange, $n=15$, were categorised as intermediate). The use of alternative cut-off criteria between the two approaches means that a difference in how an individual is categorised could reflect the difference in criteria rather than the assessment method itself. To allow for this and to further assess the influence of different cut-off criteria on the consistency of personality assignments, we also considered a series of alternative cut-offs that were uniform for both approaches. Thus, in addition to the standard (approach specific) criteria, we also categorised individuals as bold or shy that fell into (a) the upper and lower 50%, (b) the fourth and first quartiles and (c) above and below one standard deviation of the mean. Individuals that fell between these cut-offs were categorised as intermediate. We note here that the 50% cut-off and the experimental scores using the standard cut-off may not be directly comparable to other cut-offs in (2) as they do not include an intermediate category. Data were analysed in R (2.10.1: R Core Development Team 2009; package psych: Revelle 2010, package nlme: Pinhiero et al. 2009, package paran: Dinno 2009).

Results

For our observer assessments of individual boldness, the median observer ratings of boldness ranged from −2 to 2 with a median of 0.5. We calculated the intra-class correlation coefficient for rater agreement as ICC (3, k)= 0.80, indicating a high level of consistency between observers. For our experimental assessments, the first principal component of the PCA explained 50.3% of the variation in the behavioural data, while the second explained 37.3% (Table 2). Higher PC1 scores were associated with higher latencies to handle and eat the food item, and lower inspection and handling times. In contrast, lower PC2 scores were associated with longer latencies to approach and handle the novel food. In order to increase the interpretability of PC1 for these analyses, we multiplied the scores by −1. Higher values of these scores are thus indicative of bolder behaviour, i.e. individuals that were willing to spend longer time in close proximity to—and in contact with—the novel object (PC1) and to approach and interact with the novel object more quickly in the first place (PC2). They also suggest that boldness in these experiments can be broken down into two independent constituents, namely a willingness to approach and subsequently engage with unfamiliar objects (PC2 and PC1, respectively). Finally, scrutiny of the frequency distributions of the boldness scores revealed that PC1 assigned more individuals at the shyer end of the assessed axis, while the observer ratings and PC2 assigned more individuals at the bolder end of the range (Fig. 1).

We found a strong positive association between the first principal component of the experimental assessments (PC1) and the observer assessments (median observer rating) of boldness (LMM: $\beta \pm$ SE=0.30±0.07, t=4.31, df=54, P= 0.0001; Fig. 2). This relationship held despite high variation in the observer ratings for boldness in those animals that fell towards (but not on) the extremely shy end of the experimental range (i.e. scored between 0 and −0.5 on the experimental assessments: see Fig. 2). We did not find a significant effect of troop, sex or food type on this relationship (see Table 3) and thus did not include these variables in the final model. In contrast to PC1, we found no relationship between PC2 and the observer assessments (LMM: $\beta \pm$ SE=0.11±0.09, t=1.28, df=54, P=0.21). This

Table 2 Component loadings of experimental behaviours observed on the first and second principal component

Behaviour	PC1	PC2
Latency to approach food item (s)	0.316	−0.633
Latency to handle food item (s)	0.323	−0.629
Latency to eat food item (s)	0.414	0.218
Time inspecting food item (s)	−0.559	−0.276
Time handling food item (s)	−0.558	−0.283
Eigenvalue	2.516	1.864
Cumulative variance explained	50.3%	37.3%

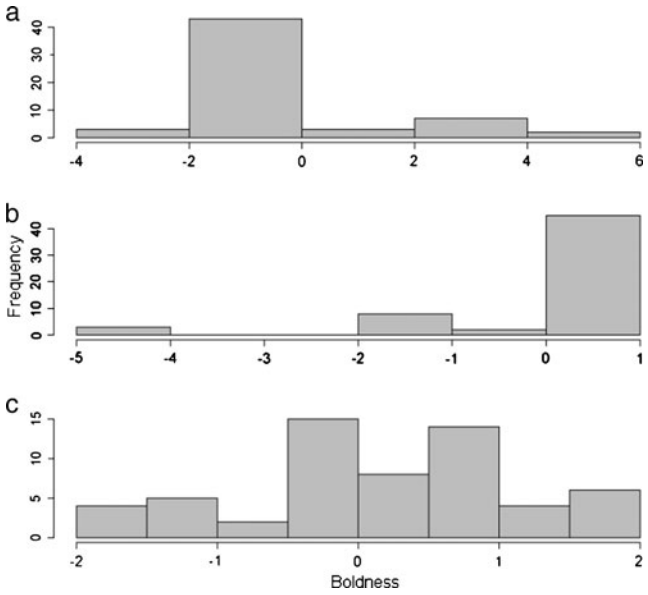


Fig. 1 Frequency distributions of the experimental scores (PC1 and PC2) of individual baboon boldness (**a** and **b**, respectively) and observer ratings of individual baboon boldness (**c**)

is unsurprising given that PC2 is, by definition, orthogonal to PC1, but it does suggest that the observer assessments of boldness better capture the 'willingness to engage' rather than 'willingness to approach' elements of boldness in the experimental assessments.

Finally, to investigate the consistency of assignments to bold and shy categories (1) and bold, shy and intermediate categories (2), we categorised each baboon as either bold, shy or intermediate based on their observer ratings and experimental PC1 scores using four categorisation criteria (Fig. 3; the PC2 scores were not used here, since the

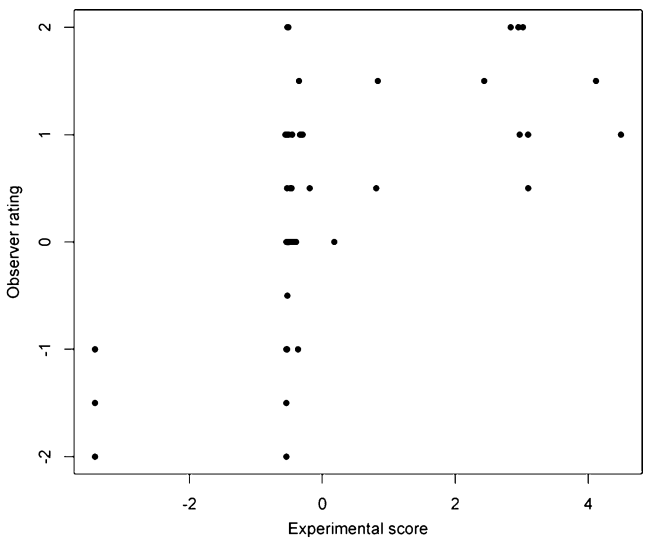


Fig. 2 Relationship between the observer rating and experimental scores of baboon boldness. Each point is one individual

Table 3 Comparisons of the fits of the null (linear) model with four linear mixed effects models, one for each random effect under investigation

Random effect	*df*	AIC	Likelihood ratio	*P* value
No effect	3	154.65		
Troop	4	156.63	0.02	0.88
Sex	4	155.40	1.25	0.26
Food type	4	155.23	1.42	0.23
Age class	4	153.17	3.48	0.06

preceding analysis found them to be unrelated to the observer ratings of boldness). For (1) the highest consistency was achieved by the 50% criterion, followed by the standard deviation, standard and quartiles criteria. For (2), after disregarding the 50% cut-off (since it involves comparisons in which there is no intermediate category), the highest consistency was achieved by the standard deviation, quartiles and standard criteria. Broadly, then, we found that more extreme cut-off criteria offered higher consistency and thus comparability, across observer ratings and experimental categorisations of individual personality. However, this conclusion refers to those cases where there are three potential categories for individual assignment (shy, bold or intermediate), even when we were only

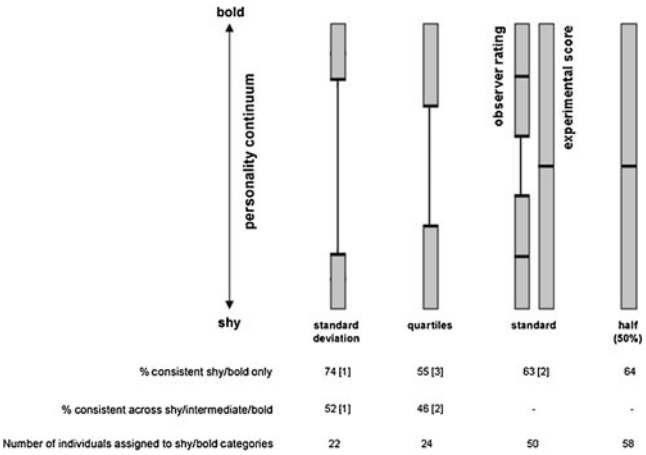


Fig. 3 Comparison of the agreement between methods using different criteria to categorise individuals as bold or shy. The standard criterion is approach specific (i.e. for observer ratings assessments, the cut-off is the midrange score on the rating scale, 0; for experimental assessments, the cut-off is the median value of PC1). For all other criteria, the cut-off is uniform for both approaches, although the exact values and hence the number of individuals involved varies between approaches depending on the data. In the diagrams, the *shaded area* indicates the area sampled in the distribution according to the relevant cut-off criterion. The figures below the diagrams indicate the percentage consistency and number of individuals assigned in each case. The *numbers in brackets* give the rank order of performance for those three cut-off criteria that are comparable across both shy/bold and shy/intermediate/bold classifications

considering shy/bold assignments. A two-way classification (where there is no intermediate category available) also performed very well, presumably reflecting the fact that there were only two groups for assignment and thus fewer opportunities for error.

Discussion

We found that the observer ratings and experimental scores of baboon boldness were positively correlated (Fig. 1). This is the first study to demonstrate such a relationship in a wild population. Nevertheless, our findings are consistent with previous studies comparing these assessment methods in captive animals. Carlstead et al. (1999), Wielebnowski (1999) and Bergvall et al. (2011) found positive correlations between rated items and experimentally tested behaviours, and Uher and Asendorpf's (2008) study on great apes found a Pearson's correlation of 0.56 between the personality scores assigned by each approach; in comparison, a Pearson's correlation of 0.56 is obtained with these study data. The similarity of these results between studies is encouraging and suggests that observer ratings and experiments of boldness are targeting the same behaviours.

One important consideration in this interpretation is that two different dimensions of boldness were identified through the principal components analysis of the experimental assessments, and only one of these showed any association to the observer ratings. Fortunately this principal component, PC1, was also the most important statistically speaking (PC1 explained over half of the variance, compared to the third of the variance explained by PC2). Thus, while our findings support a high level of congruence between observer ratings and experimental assessments, they also highlight that not all aspects of boldness captured in one approach will inevitably be captured in the other. We return to this point below. Meanwhile, for the purposes of this discussion, all further references to the experimental scores relate to PC1 only unless specified otherwise.

Although we found a strong positive relationship between the observer ratings and experimental scores, we also found both distributions to be skewed. The distribution of the experimental scores was positively skewed (i.e. scoring more individuals at the shyer end of the axis, Fig. 1). This pattern indicates that as the baboons showed an increasing willingness to interact with a novel food, the shortness of the latencies and the lengths of the inspection processing times involved provided progressively less information about how bold the individuals were (subjectively speaking). Notably, a skewed distribution, albeit reversed, is also seen in the boldness scores of Namibian rock agama *Agama planiceps* (Carter et al. 2010, see online resource 2), suggesting that such distributions may be quite widespread in experimental personality assessments. Alternatively, responses to experimental stimuli may show no tendency towards a unimodal distribution; Sinn et al. (2008) categorised individuals as bold, shy or intermediate based on a tri-modal frequency distribution of individual responses to behavioural tests (also see references in Table 1). Altogether, these findings indicate that behavioural responses to experimental stimuli should not be assumed to follow a normal distribution, and that this may complicate their interpretation.

When we categorised baboons as bold or shy according to their ratings or experimental assessments, the observer ratings did not always accurately predict whether a baboon would act in a bold or shy manner during the novel food experiment (Fig. 3). In those cases where individuals were assigned to one of three personality categories (shy, intermediate or bold) according to both the observer ratings and experimental scores, the most accurate criteria were those that only considered individuals with the most extreme ratings (the standard deviation and quartiles criteria). That is, raters could more accurately predict the reactions of the most bold or shy individuals but were less accurate at predicting the reactions of those individuals who fell nearer the middle of the bold–shy continuum. Nonetheless, there was also high overall agreement when individuals were assigned to bold or shy groups using the 50% cut-off, presumably because there were only two groups for assignment (rather than three) in this case. These results stress that there is a trade-off associated with categorising individuals as bold or shy. On the one hand, in a three-way classification, retaining a large sample size (i.e. choosing criteria that allow large numbers of individuals to be coded as bold or shy) decreases the agreement between assessment methods; the level of agreement can be increased (by choosing more extreme criteria), but this reduces the sample size. On the other hand, a two-way classification facilitates a larger sample with higher agreement between assessment methods but is disadvantaged by broader definitions of bold and shy categories (including individuals that would otherwise be termed 'intermediate') that might be counterproductive when comparing bold and shy personalities.

Although the personality scores obtained through observer ratings and experimental scores are correlated, and there can be good agreement between the two methods in their assignment of individuals to personality categories (depending on the cut-off criteria used), it is also true that the correlation between approaches is only observed with PC1 (not PC2) and the categorical assignments based on the observer ratings and PC1 are never in perfect agreement. There are two methodological differences between the ratings and experimental approaches that might account for these inconsistencies. First, the experimental

approach assesses personality relative to the sampled population, while the observer ratings approach assesses personality in absolute terms according to the definition of the items individuals are rated on (Itoh 2002). If the sample of the former does not correspond with the definitions of the latter (e.g. if only males were experimentally sampled, but the observer rating definitions were developed for both sexes), the match between the two approaches may be compromised. In the present study, this should not be a problem, since animals of all age–sex classes were experimentally sampled and the observer rating definitions were developed with the same sample in mind. Second, ratings and experimental assessments may not always correspond because of differences in specificity: the observer ratings of boldness are defined by the given questionnaire and encompass boldness in multiple contexts, whereas experimental boldness assessments are limited to the specific context of the experimental conditions, in this case encounters with novel foods (see also Nettle and Penke 2010). Thus, the two approaches may capture slightly different aspects of personality, although as we might expect—given that boldness in multiple contexts should predict boldness in a specific context—there is still a strong positive relationship between the two (at least in the case of the PC1 scores, which account for the majority of the variance in the experimental assessments).

Sih et al. (2004) and Uher (2008) highlight the divide between different approaches in animal personality research. In contrast, our analysis of ratings and experimental methods, together with those of previous captive studies, promisingly suggests that these alternative techniques of personality assessment are broadly in agreement. Two caveats arise from our study: (1) agreement between ratings and experimental assignments of personality type can depend on both the cut-off criteria and number of categories available and (2) comparisons between approaches should bear in mind subtle differences in the concepts and contexts of boldness (Itoh 2002; Sih et al. 2004; Uher 2008). Further, we would like to highlight that we are aware that not all studies of animal personality are appropriate for observer ratings either due to the sample sizes used or the species under investigation. For example, in a comprehensive review by Gosling (2001), the use of observer ratings became less frequent as the reviewed studies moved from primates through other tetrapods to bony fishes and finally insects. Nonetheless, the potential benefits of comparing and integrating across approaches are substantial, and we duly encourage those studying personality to make use of related research across disciplines (Nettle and Penke 2010) and, where appropriate, to adopt both ratings and experimental methodologies in their own research design (Vazire et al. 2007; Uher 2008; Uher and Asendorpf 2008). We also recommend that future studies should, where possible, use continuous personality scales (Nettle and Penke 2010), since these show the greatest comparability between methods. However, if individuals must be categorised, we suggest that researchers either (a) analyse only those individuals categorised as bold or shy by both ratings and experimental approaches or, if these methods cannot be employed simultaneously, (b) do not use approach-specific cut-off criteria but choose a cut-off that can be compared by both approaches. Bold/shy assignments using the 50% or standard deviation and bold/shy/intermediate assignments using the more extreme criteria, i.e. the standard deviation or quartile cut-offs, are likely to be most satisfactory. We hope that this study highlights not only the common goals adopted by those scientists using observer ratings and experimental assessments of personality, across both behavioural ecology and comparative psychology, but also the potential for collaboration between them in the future.

Acknowledgments We thank AfriKitty, Sicko, Birko, Titch, Gin-Gin, Peckers and Symie for spending enough time with the baboons to be able to assess their boldness. We are grateful to the Ministry of Lands and Resettlement for permission to work at Tsaobis Leopard Park, the Gobabeb Training and Research Centre for affiliation, and the Ministry of Environment and Tourism for research permission in Namibia. We confirm that we have adhered to the Guidelines for the Use of Animals in Behavioural Research and Teaching (*Animal Behaviour* 2006, 71:245–253). We are grateful to Jana Uher and two anonymous reviewers for comments on earlier versions of this ms. Financial support was provided by grants from the Leakey Foundation, the Animal Behavior Society (USA), the International Primatological Society, and the Explorer's Fund to AJC. AJC was supported by a Fenner School of Environment and Society scholarship and her life savings. HHM was supported by a NERC Open CASE scholarship and beer. This paper is a publication of the ZSL Institute of Zoology's *Tsaobis Baboon Project.*

References

Bell AM (2005) Behavioural differences between individuals and two populations of stickleback (*Gasterosteus aculeatus*). J Evol Biol 18:464–473

Bergvall UA, Schpäers A, Kjellander P, Weiss A (2011) Personality and foraging decisions in fallow deer *Dama dama*. Anim Behav 81:101–112

Bremner-Harrison S, Prodohl PA, Elwood RW (2004) Behavioural trait assessment as a release criterion: boldness predicts early death in a reintroduction programme of captive-bred swift fox (*Vulpes velox*). Anim Conservat 7:313–320

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56:81–105

Carere C, Drent PJ, Koolhas JM, Groothuis TGG (2005) Epigenetic effects on personality traits: early food provisioning and sibling competition. Behaviour 142:1329–1355

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

Carter AJ, Goldizen AW, Tromp SA (2010) Agamas exhibit behavioural syndromes: bolder males bask more but may suffer higher predation. Behav Ecol 21:655–661. doi:10.1093/beheco/arq036

Dinno A (2008) paran: Horn's test of principal components/factors
Freeman HD, Gosling SD (2010) Personality in nonhuman primates: a review and evaluation of past research. Am J Primatol 72:653–671
Gosling SD (2001) From mice to men: what can we learn about personality from animal research? Psych Bull 127:45–86
Horn JL (1965) A rationale and test for the number of factors In factor analysis. Psychometrika 30:179–85
Huchard E, Alvergne A, Fejan D, Knapp LA, Cowlishaw G, Raymond M (2010) More than friends? Behavioural and genetic aspects of heterosexual associations in wild chacma baboons. Behav Ecol Sociobiol 64:769–781
Itoh K (2002) Personality research with non-human primates: theoretical formulation and methods. Primates 43:249–261
Konecna M, Lhota S, Weiss A, Urbanek T, Adamova T, Pluhacek J (2008) Personality in free-ranging hanuman langur (*Semnopithecus entellus*) males: subjective ratings and recorded behavior. J Comp Psych 122:379–389
Kurvers RHJM, Prins HHT, Wieren SE, van Oers K, van Nolet BA, Ydenberg RC (2010) The effect of personality on social foraging: shy barnacle geese scrounge more. Proc R Soc Lond B 277:601–608. doi:101098/rspb20091474
McDougall PT, Réale D, Sol D, Reader SM (2006) Wildlife conservation and animal management: causes and consequences of evolutionary change for captive reintroduced and wild populations. Anim Conservat 9:39–48
McGuire MT, Raleigh MJ, Pollack DB (1994) Personality features in vervet monkeys: the effects of sex age social status and group composition. Am J Primatol 33:1–13
Nettle D, Penke L (2010) Personality: bridging the literatures from human psychology and behavioural ecology. Phil Trans R Soc B 365:4043–4050
Pederson AK, King JE, Landau VI (2005) Chimpanzee (*Pan troglodytes*) personality predicts behaviour. J Res Pers 39:534–549
Pinhiero J, Bates D, DebRoy S, Sarkar D, R Core team (2009) Nlme: linear and nonlinear mixed effects models R package version 31–96
R Development Core Team (2009) R: a language and environment for statistical computing R Foundation for Statistical Computing Vienna Austria. http://wwwR-projectorg Accessed 23 March 2010
Revelle W (2010) Psych: procedures for personality and psychological research R package version 1086
Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86:420–428
Sih A, Bell AM, Johnson JC, Ziemba RE (2004) Behavioral syndromes: an integrative overview. Quart Rev Biol 79:241–277
Sinn DL, Gosling SD, Moltschaniwskyj NA (2008) Development of shy/bold behaviour in squid: context-specific phenotypes associated with developmental plasticity. Anim Behav 75:433–442
Uher J (2008) Author's response: three methodological core issues of comparative personality research. Eur J Pers 22:475–496
Uher J, Asendorpf JB (2008) Personality assessment in Great Apes: comparing ecologically valid behavior measures behavior ratings and adjective ratings. J Res Pers 42:821–838. doi:101016/jjrp200710004
van Oers K, Klunder M, Drent PJ (2005) Context dependence of personalities: risk-taking behavior in a social and non-social situation. Behav Ecol 21:655–661. doi:101093/beheco/ari045
Vazire S, Gosling SD, Dickey AS, Schapiro SJ (2007) Measuring personality in nonhuman animals. In: Robins RW, Fraley RC, Krueger RF (eds) Handbook of research methods in personality psychology. The Guildford Press, New York, pp 190–206
Weiss A, King JE, Hopkins WD (2007) A cross-setting study of chimpanzee (*Pan troglodytes*) personality structure and development: zoological parks and Yerks National Primate Research Center. Am J Primatol 69:1264–1277
Wielebnowski NC (1999) Behavioral differences as predictors of breeding status in captive cheetahs. Zoo Biol 18:335–349