American Psychological Association

Journal of Comparative Psychology

0735-7036/19/$12.00

2019, Vol. 133, No. 3, 313–325
http://dx.doi.org/10.1037/com0000163

Measuring Personality in the Field: An *In Situ* Comparison of Personality Quantification Methods in Wild Barbary Macaques (*Macaca sylvanus*)

Patrick J. Tkaczynski, Caroline Ross, and Ann MacLarnon
University of Roehampton

Mohamed Mouna
Mohammed V University

Bonaventura Majolo
University of Lincoln

Julia Lehmann
University of Roehampton

Three popular approaches exist for quantifying personality in animals: behavioral coding in unconstrained and experimental settings and trait assessment. Both behavioral coding in an unconstrained setting and trait assessment aim to identify an overview of personality structure by reducing the behavioral repertoire of a species into broad personality dimensions, whereas experimental assays quantify personality as reactive tendencies to particular stimuli. Criticisms of these methods include that they generate personality dimensions with low levels of cross-study or cross-species comparability (behavioral coding in unconstrained and experimental settings) or that the personality dimensions generated are not ecologically valid, that is, not reflecting naturally occurring behavior (trait assessment and experimental assays). Which method is best for comparative research is currently debated, and there is presently a paucity of personality research conducted in wild subjects. In our study, all three described methods are used to quantify personality in a wild animal subject, the Barbary macaque (*Macaca sylvanus*). Our results show that the structures generated by unconstrained behavioral coding and trait assessment were not equivalent. Personality dimensions derived from both trait assessments and experimental assays demonstrated low levels of ecological validity, with very limited correlation with behaviors observed in nonmanipulated circumstances. Our results reflect the methodological differences between these quantification methods. Based on these findings and the practical considerations of wild animal research, we suggest future comparative studies of quantification methods within similar methodological frameworks to best identify methods viable for future comparisons of personality structures in wild animals.

Keywords: animal personality, quantification methods, comparative research

Supplemental materials: http://dx.doi.org/10.1037/com0000163.supp

Intraindividual consistency and interindividual variation in behavior ("personality"; Réale, Reader, Sol, McDougall, & Dingemanse, 2007) have been found in a broad range of animal taxa (Freeman & Gosling, 2010; Réale et al., 2007; Sih, Bell, & Johnson, 2004). The apparent ubiquity of personality in animals presents comparative opportunities to explore the evolutionary history of personality within and between different taxa (Adams et al., 2015; Gosling, 2008). Nonhuman animal (hereafter animal)

This article was published Online First December 27, 2018.

Patrick J. Tkaczynski, Caroline Ross, and Ann MacLarnon, Centre for Research in Evolutionary, Social and Inter-Disciplinary Anthropology, University of Roehampton; Mohamed Mouna, Institut Scientifique, Mohammed V University; Bonaventura Majolo, School of Psychology, University of Lincoln; Julia Lehmann, Centre for Research in Evolutionary, Social and Inter-Disciplinary Anthropology, University of Roehampton.

We thank Ifrane National Park, the Haut Commissariat aux Eaux et Forêts et à la Lutte Contre la Désertification, Ecole Nationale Forestiere d'Ingeniuers, and Institut Scientifique de Rabat for research permission and facilitation. We thank Liz Campbell, Melanie LaCava, Kevin Remeuf, Natalie Miller, Marin Deith, Jamie Canepa, Laetitia Maréchal, Alice Marks, Selma El Fassi-Fihri, Anna Nesbit, Alan Rincon, Anna Seltmann, and Barbora Kubovena for their contribution to collecting data and/or completing questionnaires. We thank Prof Julia Fischer and Dr Andy Radford for the Barbary macaque audio files used in the experiments. We thank Chris Herridge for his technical assistance managing the data. Finally, we thank Prof Fragaszy and three anonymous reviewers for their feedback and comments that substantially improved the clarity and quality of the manuscript.

Ann MacLarnon is now at the Department of Anthropology, Durham University.

Correspondence concerning this article should be addressed to Patrick J. Tkaczynski who is now at the Department of Primatology, Max Planck Institute of Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: patrick_tkaczynski@eva.mpg.de or pjtresearchltd@gmail.com

personality research continues to be highly descriptive and how best to quantify personality in animals remains contentious (Carter, Feeney, Marshall, Cowlishaw, & Heinsohn, 2013; Dirienzo & Montiglio, 2015; Réale et al., 2007; Uher & Visalberghi, 2016). Questions persist as to how animal personality can be measured reliably, the degree of objectivity of data collected, and whether the personality quantified reflects or is relevant to the behavioral ecology, and therefore, evolutionary history of the species (Uher, 2008; Uher & Visalberghi, 2016; Vazire, Gosling, Dickey, & Schaprio, 2007).

Although there are a wide range of methods and methodological approaches available to researchers seeking to quantify personality in animals (Carere & Maestripieri, 2013; Uher, 2008), the literature is currently dominated by three approaches: (a) "experimental assays," in which stimuli are presented to elicit personality-associated behaviors from subjects and record the frequency of their expression; (b) "unconstrained behavioral coding" (hereafter "behavioral coding"), in which behavioral observation data collected in a nonmanipulated setting are analyzed to reveal patterns of repeated behaviors and the degree to which one individual differs from another in its behavioral repertoire; and (c) "trait assessment," in which researchers familiar with individual animals complete questionnaires, rating the degree to which subjects exhibit particular personality traits or behaviors (Freeman, Gosling, & Schapiro, 2011; Uher & Asendorpf, 2008).

Experimental assays tend to quantify personality based on reactions to particular stimuli (Réale et al., 2007). For example, "boldness" is typically quantified based on responses to risky but nonnovel situations (Carere & Maestripieri, 2013; Réale et al., 2007), whereas "exploration" is typically quantified based on responses to novel situations, objects, foods or environments (Carter, Marshall, Heinsohn, & Cowlishaw, 2012a; Réale et al., 2007). Personality quantification based on reactive tendency is rooted in the "reinforcement sensitivity theory" of personality (Corr, Pickering, & Gray, 1995), which postulates that there are three components of reactivity: the "flight-or-fight system," mediating responses to aversive stimuli; the "behavioral activation system," mediating responses to positive stimuli; and the "behavioral inhibition system," mediating responses to uncertain or novel stimuli (Gray & McNaughton, 2000). As the personality dimensions of interest are chosen in advance to select an appropriate stimulus, this can make it hard to quantify personality that is comparable across species and studies, particularly if only a single experimental approach is used in a study or if stimuli used to quantify personality are highly species/population-specific (Carter et al., 2013; Uher, 2008). Furthermore, behaviors, and thus personality, elicited experimentally may not reflect "natural" behaviors and hence lack "ecological validity" (Carter et al., 2013; Réale et al., 2007), and therefore may not be relevant for studying the behavioral or evolutionary ecology of personality within or across species (Freeman, Gosling, & Schapiro, 2011; Vazire et al., 2007).

Experimental assays typically aim to identify predefined personality constructs, such as boldness or exploration, and thus use a "top-down" methodology (Uher, 2008). Trait assessments also typically use a "top-down," standardized method (using similar questionnaires for different species and studies), which facilitates phylogenetic and interstudy comparisons of the presence and absence of personality dimensions throughout taxa (Adams et al., 2015; Konečná, Weiss, Lhota, & Wallner, 2012). Much trait assessment research has utilized the human model of personality to develop questionnaires (Weiss, 2017). Although such an anthropocentric approach is contentious, it has yielded useful comparative results within the primate taxon (Adams et al., 2015; Weiss, 2017). Although less common in the literature, researchers can also develop "bottom-up" trait assessments in which the questionnaire items are derived from behaviors of the subjects (Uher & Asendorpf, 2008). Using either approach, questionnaires remain inherently subjective and based on human perceptions of personality and behavior, which makes them distinct from experimental assay and behavioral coding, which record the actual behavior of subjects. Thus, personality identified from analysis of questionnaires may not be ecologically valid and reflective of real-world behavior (Freeman et al., 2011). Furthermore, personality, by definition, is interindividual variation and intraindividual consistency in behavior (Réale et al., 2007). The presence/absence of traits for a given individual is calculated relative to the population average scores for those traits. This procedure intrinsically generates interindividual variation in ratings for certain items where the variation may actually be very low and thus not a valid component of personality. Trait assessment also requires the raters to utilize their memory of subjects; if the same raters are used to evaluate personality in subjects at different time periods, the raters are drawing upon their own mental image of the behavior of subjects, and these mental images may be more static than the actual behaviors of individuals, creating an exaggerated degree of intraindividual consistency in personality (Uher & Asendorpf, 2008; Uher & Visalberghi, 2016; Uher, Werner, & Gosselt, 2013).

Behavioral coding generally takes a "bottom-up" approach to quantifying personality (Freeman et al., 2013; Seyfarth, Silk, & Cheney, 2012) and is based on naturally occurring behavior, so derived personality dimensions, which are not constrained at the outset, have ecological validity. According to the principles of "interactionist psychology," certain situations or environments are likely to cause individuals to behave similarly, whereas other situations will result in clear interindividual differences, and thus clear differences in identified personality (Carter et al., 2013; Tett & Guterman, 2000). Therefore, when compared with "top-down" experimental assays and trait assessments, behavioral coding requires relatively intensive and time-consuming data collection to incorporate behavioral responses of individuals in a range of situations or environments.

In their review of animal personality methods, Carter et al. (2013) proposed that researchers quantifying personality must achieve three forms of validity for the personality dimensions generated: ecological validity (personality dimensions reflect naturally occurring behavior), convergent validity (correlations between the results of different methods theoretically measuring the same trait), and discriminant validity (lack of correlation between the results of different methods theoretically measuring separate traits). Experimental assays and trait assessments when used alone cannot identify ecological validity, and no method used in isolation can test for convergent or discriminant validity. Both behavioral coding and trait assessment aim to reduce the entire personality of subjects into broad dimensions and, therefore, we should expect convergent validity between dimensions derived from these two methods (Uher & Asendorpf, 2008). Experimental assays quantify personality based on reactions to particular stimuli. As such, we

should expect discriminant validity between experimental assays of different components of reactive tendencies.

Given that there are pros and cons of these popular methods, it is pertinent to establish the degree of comparability between the results derived from each method, as well as their validity. In their review, Freeman et al. (2011) highlighted several primate studies in which scores for trait assessment-derived personality factors correlated significantly with behavioral coding items when subjects were observed in nonexperimental settings. For example, in captive chimpanzees (*Pan troglodytes*), the trait assessment-derived "playfulness" correlated significantly with rates of "social play" derived by behavioral coding (Pederson, King, & Landau, 2005). However, as noted by Freeman et al. (2011), convergence between the different personality quantification methods was generally low. Indeed, recent studies of captive common marmosets (*Callithrix jacchus*; Iwanicki & Lehmann, 2015) and wild bonobos (*Pan paniscus*; Garai, Weiss, Arnaud, & Furuichi, 2016) found low levels of convergent validity, with correlation coefficients between trait assessment- and behavioral coding-derived personality expression ranging between 0.01 and 0.50 across both studies. Direct comparisons of personality quantified by trait assessments and by experimental assays have also produced varied results. In captive rhesus macaques (*Macaca mulatta*), individuals which were rated as highly "sociable" were more likely to lipsmack (a greeting behavior) in response to video playbacks (Capitanio, 1999). Similarly, in wild chacma baboons (*Papio ursinus*), "boldness" assayed using a novel food experiment correlated with "boldness" as rated by human observers (Carter, Marshall, Heinsohn, & Cowlishaw, 2012b). However, in domestic horses (*Equus caballas*; Seaman, Davidson, & Waran, 2002) and dogs (*Canis familiaris*; Kubinyi, Gosling, & Miklósi, 2015), trait assessments for subjects did not correlate with behavior in experimental assays. In the aforementioned studies where trait assessment results have been compared with either behavioral coding or experimental assay results, the items on which raters are required to rate subjects were typically adjectives, for example, "boldness" or "sociable." Studies where raters assess the degree to which subjects express specific behaviors rather than adjective-based traits have generated stronger correlations with behavioral coding or experimental assays (Uher, Addessi, & Visalberghi, 2013; Uher & Asendorpf, 2008; Uher et al., 2013).

In situ comparisons of all three of these popular approaches together have, to date, been limited to captive animal research. In tame fallow deer (*Dama dama*), "boldness" was identified by both trait assessments and an experimental assay (novel object test), and "dominance" was identified by both trait assessments and behavioral coding (Bergvall, Schäpers, Kjellander, & Weiss, 2011). Factor scores for trait-assessment "boldness" correlated significantly with experimental assay-derived "boldness," and trait-assessment "dominance" correlated significantly with behavioral coding-derived "dominance," suggesting convergent validity for these personality traits in fallow deer. More recently, Uher and Visalberghi (2016) performed a comparison of a combined behavioral method (data derived from both behavioral observation and experimental tests) with trait assessment in a study of captive capuchins (*Sapugus spp.*), finding significant differences in the personality structures quantified by the two methods, essentially arising from human bias in trait assessments.

Personality literature, especially comparisons of personality quantification methods, is mostly composed of studies conducted in captive populations (Freeman & Gosling, 2010). Individual survival in captive populations relies on adaptation to a specific environment, which may result in depletion in behavioral trait variation within a population (McDougall, Réale, Sol, & Reader, 2006). It is important to compare methods for quantifying personality in the wild to guide future research and confirm that the personality identified has ecological validity in a setting most reflective of the species' evolutionary and ecological history.

Therefore, in our study, we incorporated and compared the three most common approaches to quantify personality in a wild primate species, Barbary macaques (*Macaca sylvanus*). In particular, we aimed to clarify how comparable the methods are (based on calculations of convergent and discriminant validity), whether each method generates dimensions that are ecologically valid, that is, reflective of natural behavior rather than solely perceptions of behavior (trait assessments) or artificially induced behavior (experimental assays), and, finally, we considered the practicality of these methods for use with wild animal subjects.

To date, there are only two published studies pertaining to Barbary macaque personality, both of which employed a "top-down" trait assessment method (Adams et al., 2015; Konečná et al., 2012). Konečná et al. (2012) studied a group of semi-free-ranging macaques in Gibraltar ($n = 27$) and described a personality structure containing four components: "Friendliness," "Activity/Excitability," "Confidence," and "Opportunism." Adams et al. (2015) studied two wild groups of macaques in Morocco ($n = 74$) and also found a four-component structure to the personality of subjects: "Friendliness," "Confidence" (both also found in Konečná et al., 2012), "Openness," and "Irritability." Adams et al. (2015) equated "Activity/Excitability" found in the previous study with the "Openness" found in their study, and "Opportunism" and "Irritability" were found to share a number of constituent traits. Nevertheless, despite using the same questionnaire, differences in the personality structure of Barbary macaques were found between the studies.

Method

All data collection was conducted following ethical approval by University of Roehampton and the receipt of research permits for the field work were provided by Haut-Commissariat aux Eaux et Forêts et à la Lutte Contre la Désertification, Royaume du Maroc.

Study Subjects/Site

Data were collected at a study site in the oak and cedar forest within the Ifrane National Park, Morocco (33° 24′ N, 05° 12′ W; elevation 1,500–2,000 m above sea level). For this study, the adults of two groups of fully habituated Barbary macaques were the subjects. Adults were defined as sexually mature individuals based on body size in both sexes, the presence of anogenital swellings during the breeding season in females, and descended testicles and large canines in males (Fooden, 2007). There were 12 subjects in the "Blue" group (five males; seven females) and 15 in the "Green" group (seven males; eight females).

Data Collection

Behavioral data. For this study, data on activity state (feeding, resting, traveling, grooming) were recorded continuously using focal animal sampling. Contact between individuals, proximity, agonistic, solitary and sexual behaviors, as well as facial displays and vocalizations were recorded as point events (Table S1 in the online supplementary materials lists the variables and provides brief definitions). To measure gregariousness, at the start and end of each focal sample the number of group members within 0–1 m, 1–5 m, and 5–10 m of the focal subject was also recorded. Finally, during these measures of proximity, observers also recorded whether subjects were central or peripheral within the group in terms of spatial position; individuals were considered peripheral if they were the outermost individual in either the front, rear or side of the group.

Behavioral data were collected in three time blocks: mating season, starting with first observed copulation and finishing with last observed copulation in the groups (11th October 2013–20th January 2014 for the Blue group; 9th October 2013–9th January 2014 for the Green group); nonmating season 1 (21st January 2014–5th March 2014 for the Blue group; 10th January–6th March 2014 for the Green group); and nonmating season 2 (4th February 2015–18th April 2015 for both groups). Division of behavioral data into time blocks allowed us to test for repeatability of behavioral variables across both time and varying social and ecological contexts (see below). Overall, 28 behavioral variables (21 behaviors and seven indices [grooming and proximity]; Table S1 in online supplementary materials) relevant to Barbary macaque socioecology (Hodges & Cortes, 2006; Koski, 2011; Neumann, Agil, Widdig, & Engelhardt, 2013) were collected/calculated. Behavioral variables were standardized before analysis using z scores.

Behavioral data were collected through focal (30 min) sampling and proximity scans at the start and end of focal samples (Altmann, 1974) using a Psion handheld computer and The Observer XT software Version 8.0. The order of subjects for focal samples was randomized; subjects were not resampled until all other individuals had been sampled and never more than once on the same day. Including the principal investigator (PJT), seven people collected behavioral data. Interobserver reliability tests using intraclass coefficients (ICC; Shrout & Fleiss, 1979) were conducted for all researchers: Researchers collected behavioral data only once they had collected two consecutive focal follows where the frequency and duration of variables recorded were significantly reliable (ICC $>$ 0.95; $p <$ .05) compared with those recorded by PJT.

A total of 1,308 hr of focal samples were collected, comprising an average of 48.41($\pm$1.64) hr per subject. A total of 5,352 proximity scan samples were recorded, comprising an average of 198.22($\pm$5.64) scans per subject.

Questionnaire data. Questionnaire data were collected for the trait assessment method of quantifying personality. Eight researchers not involved in behavioral data collection for the current project, who each had a minimum of 3 months of experience observing the study subjects, completed a questionnaire assessing the personality structure of the subjects. The questionnaire, which was used in a previous Barbary macaque personality study (Konečná et al., 2012), was adapted from the Chimpanzee Personality Questionnaire (King & Figueredo, 1997), which itself was derived from questionnaires used in human personality research (Konečná et al., 2012). The questionnaire consisted of 51 items (personality traits; see Table S2 in online supplementary materials), which were rated on a 7-point scale. A score of "1" suggests the rater believes the trait is absent in the individual, whereas a score of "7" implies the rater believes the individual exhibited "extreme amounts" of the trait. Questionnaire data were collected from researchers who had worked with the subjects between January 2012–December 2013 (n = 4), January 2013–August 2013 (n = 2), and September 2013–April 2014 (n = 2).

Experimental data. The two experimental assays of personality (boldness and exploration) were conducted between October 2013–March 2014 and October 2014–March 2015. For the experimental "boldness" assay (positive or nonfearful reactions to a threatening but nonnovel stimulus; Réale et al., 2007), playback experiments were conducted. Subjects were presented with four playbacks (three treatments and one control) over the study period. The treatment stimulus was designed to simulate intergroup encounters and consisted of aggression growls and alarm barks from nongroup conspecifics. The control stimulus was a brown-necked raven (*Corvus ruficollis*) call, a common and frequently heard bird at the field site.

Stimuli were broadcasted using an SME-AFS Portable Field Speaker (Saul Mineroff Electronics). The speaker was hidden using branches and placed ~30–50 m away from the group. Immediately following the playing of the stimulus, a 30-min focal observation was carried out with predetermined subjects (from a randomized order; one to three subjects per experiment). Researchers recorded the duration or frequency of behaviors associated with elevated anxiety or stress coping mechanisms in Barbary macaques and other primates (Schino, Perretta, Taglioni, Monaco, & Troisi, 1996; Semple, Harrison, & Lehmann, 2013), namely, self-grooming (duration), self-scratching, yawns, and vigilance (counts; vigilance was defined as sustained gazes away from general activity).

For the "exploration" experimental assay (positive or nonfearful reactions to a novel stimulus; Réale et al., 2007), novel object experiments were conducted. For treatment experiments (mean number per subject = 4 $\pm$ 2), subjects were presented with similarly sized, brightly colored toys or household items (see Figure S1 in online supplementary materials for photographs of all objects). For control experiments (mean number per subject = 2 $\pm$ 1), bundles of fallen branches were presented. To present both treatments and controls, the objects were tied to a brown (partially camouflaged) rope and suspended from a tree, ~0.5 m above the ground. Experiments were set up in advance and out of sight of the approaching group. When the group came within 30 m of the stimulus, the object was raised briefly by the researcher to draw attention to the object and then left suspended above the ground. Data collection began when the first group member (including infants and juveniles) approached within 20 m of the object. A 30-min observation of the object was conducted; the researcher recorded all individuals entering or leaving the 20-m zone around the object. All instances of interactions (visually inspecting, handling) with the object were recorded.

Quantifying Personality Dimensions

Behavioral coding. Quantifying personality using behavioral coding involves two key steps: first, identifying within subjects the behaviors that were expressed in a repeatable manner (similar frequency or duration over time and context); second, reducing the number of repeatable behaviors into broad personality dimensions (Freeman et al., 2011; Koski, 2011). To examine whether subjects expressed these variables consistently across the three time periods, each variable was examined using an analysis of variance-based measure of repeatability (R_A; Nakagawa & Schielzeth, 2010). This method calculates an estimate of repeatability, R_A, for each behavioral variable based on the degree of within-individual variation for the expression of each variable compared with the overall between-individual variation in expression for the same variable (Nakagawa & Schielzeth, 2010). Repeatability was calculated using the *rptr* package in R (Nakagawa & Schielzeth, 2010).

Behavioral variables which showed significant repeatability across the three time periods were then used in a factor analysis (FA) to identify suites of correlated variables, with mean values for each variable for each subject included in the analysis (Budaev, 2010; Koski, 2011). As the ratio of subjects to variables was low, we employed regularized FA, a method recommended for low sample number FA (Jung & Lee, 2011) specifying unweighted least squares for factor extraction. Parallel analysis was used to determine eigenvalues and number of factors (Horn, 1965) using the *paran* (Dinno, 2012) package in R. The suitability of the data set for FA was assessed using the Kaiser Meyer Olkin (KMO) measure of sampling adequacy and Barlett's test of sphericity, both performed using the *psych* package (Revelle, 2018) in R. An oblique ("Promax") rotation was used, which allows for correlations between factors, as no assumption is made about the underlying structure of data (Budaev, 2010; Koski, 2011; Neumann et al., 2013). The sum of the mean values of the behavioral variables which loaded onto a particular personality dimension was used to create individual (subject specific) scores for a particular personality dimension. Salient loadings of items on factor or components were defined as $\geq \pm 0.40$ (Konečná et al., 2012).

Trait assessment. Trait assessment analysis involves two key steps: first, determining the reliability of the ratings by examining correlations between raters for a given subject and given personality trait; second, identifying correlations among these reliably rated traits to create broad personality dimensions. Using the questionnaire data, interrater reliability for item ratings was calculated using ICCs (Shrout & Fleiss, 1979). Two coefficient types were used: ICC (3, 1), which indicates the reliability of individual ratings from one trait to another; and ICC (3, *k*), which indicates the reliability of individual ratings of a trait to a mean score of this trait based on *k* raters (Shrout & Fleiss, 1979). ICC calculations were conducted using IBM SPSS Statistics 21 (IBM Corp, Armonk, NY).

Interrater reliability for an item was defined as $p < .05$ for both ICCs, ICC[3, *k*] $\geq$ 0.75 (DeVon, Block, Moyle-Wright, Ernst, Hayden, Lazzara, Savoy, & Kostas-Polston, 2007; DeVon et al., 2007). For items with significant interrater reliability, mean rating values were calculated for each trait for each subject. Following the same FA protocol used for behavioral coding, regularized FA was then applied to reliably rated items to identify correlated items. Once factors were identified, scores for subjects for a particular personality dimension were created by summing the mean ratings for items which loaded (salient loading $\pm < 0.40$) onto a particular factor.

Experimental assays. For the Boldness assay, expressions of the behavioral variables during an experiment were converted into proportions of observation time for self-grooming and into frequencies per minute for point behaviors (self-scratch, yawn, and vigilance). These variables were standardized and then summed into a single value and multiplied by -1, so that higher scores indicated less fearful and thus bolder responses to the stimuli, to create an index of boldness. For the Exploration assay, the frequencies per minute were calculated for each type of interaction with the stimuli, with observation time being defined as the amount of time the subject was within 20 m (and thus visible range) of the object. These values were then summed to create an index of exploration.

Linear mixed-effects models were used to compare indices of boldness and exploration derived from treatment experiments to those derived from control experiments to ensure that the indices reflect responses to either a nonnovel, risky stimulus (in the case of playbacks) or a novel stimulus (in the case of novel object presentations), rather than general responses to stimuli. For this analysis, index scores for each individual were the dependent variable, individuals were included as random effects, and experiment type as a fixed effect; significance ($p < .05$) was determined by *F* tests of the fitted full model (Whittingham, Stephens, Bradbury, & Freckleton, 2006). Models were fitted using the *nlme* package in R (Pinheiro, Bates, DebRoy, Sarkar & R Core Team, 2016). Repeatability across treatment experiments of boldness and exploration measures were analyzed using the analysis of variance-based measure of repeatability (R_A; Nakagawa & Schielzeth, 2010), using the *rptr* package in R (Nakagawa & Schielzeth, 2010).

Validation of Personality Dimensions

Percentage bend correlations (Wilcox, 1994) were used to test for convergent and discriminant validity among the personality dimensions generated by behavioral coding, trait assessment, and experimental assays using standardized (*z* scores) mean values for personality dimension scores. Ecological validities of trait assessment- and experimental assay-derived dimensions were determined by examining percentage bend correlations between mean personality scores derived through each of these methods and mean expression of the behavioral coding variables (Table S1 in the online supplemental materials). Percentage bend correlations were calculated using *WRS* package in R (Wilcox & Schönbrodt, 2009).

To determine whether individuals differed significantly in their expression of the personality dimensions identified, linear mixed-effects models were used with individuals as random effects, sex and group as fixed effects, and personality scores as the dependent variable. Significance of the fixed effects was determined by *F* tests of the fitted full model (Whittingham et al., 2006). Likelihood ratio tests compared models with and without random effects to determine if there was significant interindividual variation in personality scores. Models were fitted using the *nlme* package in R (Pinheiro et al., 2016).

Results

Personality Dimensions

Behavioral coding. Of the 28 behavior variables included in the behavioral coding approach, 18 were repeatable (see Table S3 in online supplementary materials for all R_A coefficients). The KMO measure of sampling adequacy (0.61) and Bartlett's test of sphericity ($\chi^2 = 839.74$, $df = 153$; $p < .01$) confirmed the suitability of these repeatable variables for FA. Parallel analysis of behavioral variables suggested three components to be extracted. Three variables were removed following the first FA due to insufficient loading (≤0.40) on any component (Activity [maximum loading = −0.35], Retreat [−0.21], Body Shake [−0.21]). Parallel analysis of the new data set suggested the extraction of three factors; the KMO measure of sampling adequacy (0.69) and Bartlett's test of sphericity ($\chi^2 = 766.32$, $df = 105$; $p < .01$) confirmed the suitability of this data set for FA.

The final FA generated three factors that explained 56% of total variance (see Table 1); each factor from behavioral coding is indicated by a $_{BC}$ suffix after the name. The first factor (accounting for 20% of variance) had positive loadings for eight variables related to dominance, prosocial behaviors, as well as behaviors associated with anxiety. This dimension was called "Excitability$_{BC}$," a dimension of personality previously identified in Barbary macaques using trait assessment (Konečná et al., 2012). The second factor (accounting for 19% of variance) had positive loadings for centrality and number of neighbors within 5 m, so was called "Sociability$_{BC}$," a term used in previous macaque personality research (Neumann et al., 2013; Sussman, Ha, Bentson, & Crockett, 2013). The third factor (accounting for 17% of variance) contains four variables related exclusively to grooming, either self- or allogrooming. Previous personality studies have related factors containing grooming variables to sociability (Neumann et al., 2013). However, the high loading for self-grooming, a solitary activity, found here suggests that the factor identified is independent of sociability. Therefore, the term "Tactility$_{BC}$" was introduced.

Table 1
Loadings of Behavioral Variables From Factor Analysis Used in Behavioral Coding Method

Variable	Factor 1: Excitability$_{BC}$	Factor 2: Sociability$_{BC}$	Factor 3: Tactility$_{BC}$
Triadic embrace	**.76**	.00	.10
Yawn	**.74**	−.21	−.31
Embrace	**.73**	.10	−.15
Tree shake	**.73**	−.33	.13
Open mouth	**.64**	.19	−.17
Mount	**.62**	.12	−.04
Genital touch	**.61**	.10	.37
Contact aggression	**.51**	.17	.37
Central	.01	**.91**	.21
Peripheral	−.02	**−.91**	−.21
Neighbors within 5–10 m	−.10	**.85**	−.17
Allogroom	−.20	.19	**.84**
Neighbors within 1 m	.18	.31	**.71**
Grooming density	−.31	−.05	**.66**
Self-groom	.04	−.31	**.60**

Note. BC = behavioral coding. Salient loadings (>.40) are in bold. Variables that loaded significantly on more than one factor are in bold and italicized. Higher loadings determined which factor a variable was included in.

Table 2
Intraclass Coefficient (ICC) Values for Questionnaire Items

Item	ICC (3, 1)	ICC (3, *k*)	Item variance	*F*	*p* value
Dominance	.59	.91	0.57	10.92	<.01
Eccentric	.63	.91	2.57	13.39	<.01
Irritability	.54	.89	1.26	9.39	<.01
Submissive	.50	.88	3.61	7.80	<.01
Solitary	.50	.87	1.61	8.50	<.01
Popular	.47	.86	1.34	10.10	<.01
Timid	.46	.85	1.70	6.63	<.01
Equable	.46	.85	0.57	6.78	<.01
Depressed	.44	.85	2.97	6.59	<.01
Insecure	.43	.84	4.74	6.47	<.01
Independent	.42	.83	1.50	5.91	<.01
Disorganized	.45	.83	1.21	6.96	<.01
Sociable	.41	.83	1.62	7.39	<.01
Fearful	.41	.83	1.40	5.86	<.01
Tense	.40	.82	2.55	6.85	<.01
Protective	.38	.81	0.48	5.52	<.01
Helpful	.36	.80	0.34	6.95	<.01
Erratic	.36	.80	2.06	5.45	<.01
Aggressive	.35	.79	1.94	5.52	<.01
Confidence	.34	.78	1.41	4.71	<.01
Gentle	.31	.76	1.27	4.67	<.01
Affectionate	.30	.75	0.71	5.13	<.01
Excitable*	.28	.73	0.18	3.73	<.01
Intelligence*	.27	.73	1.40	4.39	<.01
Consistent*	.28	.73	0.76	4.27	<.01
Impulsive*	.27	.72	1.75	3.78	<.01
Friendly*	.27	.72	0.76	4.72	<.01
Manipulative*	.27	.72	1.60	3.98	<.01
Playful*	.27	.72	0.30	5.50	<.01
Persistent*	.26	.71	0.52	3.60	<.01
Sympathetic*	.25	.70	0.57	5.21	<.01
Socially playful*	.24	.69	0.88	4.81	<.01
Permissive*	.24	.69	1.33	3.30	<.01
Conventional*	.23	.67	0.78	3.68	<.01
Bullying*	.22	.66	1.75	3.17	<.01
Patient*	.20	.64	0.30	2.92	<.01
Sensitive*	.19	.62	0.79	3.46	<.01
Unemotional*	.18	.60	0.26	2.50	<.01
Active*	.18	.60	0.59	3.46	<.01
Selective*	.17	.60	1.94	2.57	.01
Jealous*	.16	.58	1.19	2.29	.02
Assertive*	.15	.54	0.54	2.23	.03
Cautious*	.13	.52	1.43	2.06	.04
Reckless*	.13	.51	1.40	2.03	.05
Stingy*	.02	.07	3.29	1.08	.38
Lazy*	.06	.21	0.13	1.33	.16
Explorative*	−.06	−.27	6.41	0.76	.79
Alert*	.00	.00	4.02	1.00	.48
Curious*	−.04	−.19	3.84	0.80	.75
Opportunistic*	−.01	−.05	3.46	0.93	.57
Inventive*	−.05	−.25	0.32	0.71	.85

Note. Items marked with asterisks were not significantly reliably rated ($p > .05$).

Trait assessment. Seven of the 29 questionnaire items were unreliably rated by raters (see Table 2). The ICC(3,1) coefficients for the remaining 22 items ranged from 0.30 to 0.63 with a mean of 0.43 (±0.083); ICC(3, k) coefficients for these items ranged from 0.75 to 0.91 with a mean of 0.83 (±0.04; Table 2). Parallel analysis of ratings suggested the data should be reduced to three

Table 3
Loadings of Questionnaire Items From Principal Component Analysis Used in Trait Assessment Method

Variable	Component 1: Confidence$_{TA}$	Component 2: Friendliness$_{TA}$	Component 3: Neuroticism$_{TA}$
Aggressive	**.92**	−.38	.23
Dominant	**.91**	.13	−.07
Insecure	**−.91**	.12	.22
Timid	**−.90**	−.01	.15
Confident	**.87**	.23	.11
Submissive	**−.83**	.28	.38
Fearful	**−.79**	−.05	.33
Independent	**.74**	.06	***.51***
Irritable	**.73**	***−.48***	.35
Popular	**.67**	**.55**	−.02
Protective	**.59**	***.46***	.23
Gentle	−.35	**.94**	.31
Helpful	−.04	**.92**	.04
Affectionate	.01	**.89**	.14
Equable	.05	**.77**	.08
Sociable	.25	**.67**	−.08
Eccentric	.06	.31	**.91**
Depressed	−.16	.16	**.91**
Solitary	−.13	.01	**.66**
Disorganized	.02	.02	**.64**
Tense	***.43***	−.29	**.52**
Erratic	.12	−.29	**.41**

Note. TA = trait assessment. Salient loadings (>.40) are in bold. Items that loaded significantly on more than one component are in bold and italicized. Higher loadings determined which component a variable was included in.

factors; the KMO measure of sampling adequacy (0.59) and Bartlett's test of sphericity ($\chi^2 = 743.03$, $df = 231$; $p < .01$) confirmed the suitability of this data set for FA. The three factors accounted for 74% of variation within the data set (see Table 3); each factor from trait assessment is indicated by a $_{TA}$ suffix.

The first component (accounting for 35% of total variance) contained positive loadings for items such as dominant, confident, and aggressive and was therefore called "Confidence$_{TA}$," a component found in previous Barbary macaque research (Adams et al., 2015; Konečná et al., 2012). The second component (22% of total variance) that had positive item loadings for sociable, affectionate, and gentle, was similar in structure to "Friendliness" found in previous Barbary macaque research (Adams et al., 2015; Konečná et al., 2012) and thus was called "Friendliness$_{TA}$." The final component (17% of total variance) contained positive loadings for items such as eccentric, disorganized, and solitary and was thus similar in structure to human Neuroticism (Weiss, 2017); thus, the term Neuroticism$_{TA}$ was applied to this component.

Experimental assays. Personality quantified using experimental assays in this study is denoted by the EA suffix. For the boldness assay, subjects were presented with three playback treatment experiments and one control experiment. For the exploration assay, subjects participated (entered into proximity with the object) on average in 4.2 (±1.2) treatment experiments and on average 2.40 (±1.26) control experiments. Subjects demonstrated significantly repeatable "Boldness$_{EA}$" scores across treatment experiments ($R_A = 0.20$; $SE = 0.13$; $p < .05$); mean "Boldness$_{EA}$" scores were significantly higher in treatment experiments compared with control experiments ($F = 15.95$; $df = 1, 26$; $p < .01$). Subjects did not demonstrate significantly repeatable "Exploration$_{EA}$" scores across treatment experiments ($R_A = 0.03$; $SE = 0.01$; $p = .67$); mean "Exploration$_{EA}$" scores were significantly higher in treatment experiments compared with control experiments ($F = 8.49$; $df = 1, 26$; $p = .01$).

Validation of Personality Dimensions

Convergent and discriminant validity. Only three significant correlations between the eight personality variables identified were observed (see Table 4): Positive correlations were observed between Sociability$_{BC}$ and Friendliness$_{TA}$ scores and between Sociability$_{BC}$ and Confidence$_{TA}$ scores. A positive correlation was also observed between Boldness$_{EA}$ and Tactility$_{BC}$ scores. Thus, contrary to expectations, levels of convergent validity between behavioral coding- and trait assessment-derived personality dimensions were low. Discriminant validity (a lack of a correlation) was observed between mean Boldness$_{EA}$ and Exploration$_{EA}$ scores, as expected.

Ecological validity. Mean scores for trait assessment- and experimental assay-derived dimensions rarely correlated with behavioral variable values, suggesting low levels of ecological validity (see Table 5). For trait-assessment dimensions, 16.7% of mean behavioral variable values correlated with mean personality construct scores. These correlations were consistent with the definitions of the constructs, for example, mean Friendliness$_{TA}$ scores correlated positively with average time spent centrally and average number of neighbors within 5 m. Mean Boldness$_{EA}$ scores correlated with only two of 28 mean behavioral variable values: activity and vigilance; mean Exploration$_{EA}$ scores did not correlate with any of the 28 mean behavioral variable values.

Table 4
Percentage Bend Correlation Coefficients (ρ) Between Subject Mean Scores for All Personality Dimensions (n = 27)

	Excitability$_{BC}$	Sociability$_{BC}$	Tactility$_{BC}$	Boldness$_{EA}$	Exploration$_{EA}$
Confidence$_{TA}$	.36[a]	**.45**[a]	−.29[a]	.04	−.20
Friendliness$_{TA}$	−.13[a]	**.61**[a]	.09[a]	.16	−.14
Neuroticism$_{TA}$	.22[a]	−.34[a]	−.13[a]	−.18	.30
Boldness$_{EA}$	.07	.29	**.47**	—	−.17
Exploration$_{EA}$	.26	−.24	.05	−.17	—

Note. BC = behavioral coding; EA = experimental assays; TA = trait assessment. Significant correlations indicative of convergent validity are in bold ($p < .05$).
[a] Values around which convergent validity was predicted.

Table 5
Percentage Bend Correlation Coefficients (ρ) Between Subject Means for Behavioral Variables and Personality Dimension Scores (n = 27)*

Behavior variable	$Confidence_{TA}$	$Friendliness_{TA}$	$Neuroticism_{TA}$	$Boldness_{EA}$	$Exploration_{EA}$
Activity	−.21	.01	−.36	**.41**	−.02
Submissions	−.23	−.22	−.04	.25	.06
Retreats	−.24	.01	.22	−.09	−.02
Supplants	.22	−.09	−.23	−.03	−.09
Self-groom	−.21	−.19	.07	−.02	.24
Self-scratch	−.15	−.17	**.58**	−.30	−.10
Body shake	.04	.04	.09	−.01	.07
Yawn	.21	−.04	.23	−.24	−.12
Tree shake	.26	−.20	.22	−.16	−.10
Mounting	.17	−.08	.09	.06	.01
Allogrooming	−.31	.11	−.19	.33	.23
Grooming density	−.26	.03	−.17	.34	.36
Grooming diversity	−.29	−.14	−.05	.35	.21
Grooming evenness	−.33	−.20	.02	.34	.25
Contact aggression	.25	−.13	−.10	−.14	−.04
Noncontact aggression	.29	−.31	.06	−.09	−.09
Open mouth	.24	−.27	.10	−.12	−.02
Bare teeth	.37	−.17	.01	−.27	−.12
Teeth chatter	.32	−.32	.12	−.26	−.14
Lip smack	.09	.03	.19	.20	−.04
Embrace	.35	.18	.01	−.22	.10
Genital touch	.06	.03	**.46**	−.16	.36
Sandwich	.21	−.12	.18	−.04	−.01
Vigilance	.01	−.11	.36	**.41**	−.13
Peripheral	−.15	**−.38**	.19	.05	−.17
Central	.14	**.38**	−.19	−.05	.17
Neighbors within 0–1 m	−.05	.07	−.29	.14	.15
Neighbors within 1–5 m	.30	**.38**	−.19	−.10	.15
Neighbors within 5–10 m	.21	.33	−.21	.01	.05
Approaches	.20	.07	.14	−.34	.13

Note. TA = trait assessment; EA = experimental assays. Significant correlations indicating functional validity are in bold ($p < .05$).
* Frequency per minute, proportion of observation time or index; See Table S1 in the Online Supplementary Materials.

Interindividual variation. Interindividual variation was observed in scores for six of the eight personality dimensions tests. There was no evidence of interindividual variation in scores for $Boldness_{EA}$ and $Exploration_{EA}$ (see Table 6).

Discussion

This study presents an in situ comparison of the three predominant methods used for personality quantification in *wild* animals. When testing the ecological validity and degree of convergent/discriminant validity between the methods, the results suggest that these approaches are not equivalent to one another and highlight their methodological differences.

Barbary Macaque Personality Structure

Six out of the eight quantified dimensions conformed to the definition of personality in demonstrating both interindividual variation and intraindividual consistency (Réale et al., 2007): $Excitability_{BC}$, $Sociability_{BC}$, $Tactility_{BC}$, $Confidence_{TA}$, $Friendliness_{TA}$, and $Neuroticism_{TA}$. Based on their observed convergent validity, these dimensions can be further reduced to Confidence, General Sociability ($Sociability_{BC}$ and $Friendliness_{TA}$), Tactile Sociability, and Neuroticism. This structure bears similarity to that which was previously quantified using trait assessment alone for a population of semi-free-ranging Barbary macaques (Konečná et al., 2012), that consisted of Excitability, Friendliness, Confidence, as well as an additional dimension, Opportunism. In the semi-free-ranging population, where interactions with human tourists and food provisioning are frequent (Maréchal, MacLarnon, Majolo, & Semple, 2016), Opportunism was characterized by high ratings for "Manipulative," "Jealous," or "Bullying" items on the questionnaire. Provisioning of food by tourists may encourage the expression of these traits, and hence the prominence of the combined dimension, Opportunism. Future comparative intraspecific research may further inform how personality develops differently or is differently expressed in different environments.

In our study, Sociability was split into two dimensions (General and Tactile), whereas in the semi-free-ranging population, only Friendliness was found. It has been argued that the Friendliness dimension of macaques is a "blended" dimension, incorporating elements of Extraversion and Agreeableness, which are typically separate dimensions in great apes (Adams et al., 2015). Our results suggest that combining multiple methods to quantify personality can reveal subtleties in personality structure which may be lost when using a single approach in isolation.

Table 6
Results From Linear Mixed-Effect Models Explaining Interindividual Variation in the Expression of Identified Personality Dimensions

Models	Estimate	*SE*	*F*	*p* value	Random effects	
					LRT	*p* value
$Excitability_{BC}$						
Intercept	*1.32*	*0.19*	*6.61*	*<.01*	*4.80*	*.03*
Group	*−0.73*	*0.22*	*−3.36*	*<.01*		
Sex	*−1.72*	*0.21*	*8.03*	*<.01*		
$Sociability_{BC}$						
Intercept	*1.13*	*0.07*	*527.58*	*<.01*	*14.51*	*<.01*
Group	0.07	0.08	0.39	.54		
Sex	0.03	0.08	0.08	.78		
$Tactility_{BC}$						
Intercept	*0.47*	*0.04*	*452.60*	*<.01*	*9.53*	*<.01*
Group	<−0.01	0.04	0.10	.75		
Sex	*0.27*	*0.04*	*22.57*	*<.01*		
$Confidence_{TA}$						
Intercept	*40.18*	*1.54*	*26.14*	*<.01*	*39.55*	*<.01*
Group	*−11.25*	*1.76*	*−6.40*	*<.01*		
Sex	2.29	1.73	1.32	.19		
$Friendliness_{TA}$						
Intercept	*21.94*	*1.29*	*17.01*	*<.01*	*20.45*	*<.01*
Group	−0.33	1.45	−0.22	.82		
Sex	−0.24	1.42	−0.17	.87		
$Neuroticism_{TA}$						
Intercept	*17.59*	*1.57*	*11.21*	*<.01*	*44.709*	*<.01*
Group	−2.41	1.79	−1.34	.18		
Sex	3.16	1.77	1.79	.07		
$Boldness_{EA}$						
Intercept	*0.87*	*0.04*	*1,716.91*	*<.01*	0.65	.42
Group	0.07	0.04	2.23	.15		
Sex	0.01	0.04	0.05	.83		
$Exploration_{EA}$						
Intercept	*0.60*	*0.54*	*12.80*	*<.01*	<0.01	>.99
Group	−0.31	0.24	1.68	.21		
Sex	0.19	0.24	0.60	.45		

Note. LRT = likelihood ratio tests; BC = behavioral coding; TA = trait assessment; EA = experimental assays. Significant effects ($p < .05$) are italicized; for group, Blue group is the reference factor, for sex, males are the reference factor.

The Excitability dimension identified in our study of Barbary macaques appears structurally similar to the Shy-Bold/Proactive-Reactive axis often explored in experimental assays (Coppens, de Boer, & Koolhaas, 2010; Koolhaas et al., 1999). Macaques with higher "Excitability" scores were more active and engaged in more frequent brief social interactions (either affiliative or agonistic), which are traits also characteristic of Bold/Proactive individuals (Coppens et al., 2010; Koolhaas et al., 1999). It is unfortunate that the Boldness assay used in our study failed to meet the criteria of personality (intraindividual consistency and interindividual variation) to better explore the relationships between Boldness and Excitability in Barbary macaques.

Comparing Methods Based on Validity and Practicality in the Wild

As outlined previously, trait assessment in animals, particularly using a "top-down" approach derived from human personality research, faces criticism of anthropomorphism, generating interindividual differences where they may not exist and exaggerating the stability of personality dimensions over time and context. Testing temporal consistency of personality derived from trait assessment can be achieved if the assessments are conducted by different individuals working with the subjects at different time periods, as was done in our study, so that reliability between raters for a given trait for a subject reflects consistency over time as well. Finding enough raters for this may be problematic in wild animal research, particularly for field sites where researchers of several nationalities work together and may have markedly different interpretations of the definitions of items in questionnaires (Uher & Visalberghi, 2016). Furthermore, there appears to be a discrepancy in the criteria set for interrater reliability between studies conducted in humans and in animals. In human psychometric research, reliability coefficients of at least 0.70 are considered acceptable (reviewed in DeVon et al., 2007). Criteria for interrater reliability in animal research are typically lower and highly variable between studies (e.g., ≥0.00 in Bergvall et al., 2011; ≥0.60 in Iwanicki & Lehmann, 2015).This may reflect the difficulty of applying adjectives derived from human personality to animal subjects. Questionnaires can instead be built based on species-specific behaviors, and several primate studies using this form of trait assessment gener-

ated results more reflective of actual behavior (Uher, 2008; Uher, Adessi, et al., 2013; Uher & Asendorpf, 2008; Uher & Visalberghi, 2016; Uher, Werner, et al., 2013), suggesting this approach could improve the standard of interrater reliability within the field of animal personality.

Both behavioral coding and trait assessment aim to identify a summary of personality structure among subjects or even for a species and thus convergent validity was expected between dimensions identified by these methods. As highlighted above, positive correlations were identified between the scores for trait assessment- and behavioral coding-derived dimensions, for example, $\text{Sociability}_{\text{BC}}$ and $\text{Friendliness}_{\text{TA}}$. In our study and others (Garai et al., 2016; Iwanicki & Lehmann, 2015), personality dimensions unique to a method, that is, having no correlation with dimensions from other methods, were created by trait assessment. In our study, Neuroticism was only found using trait assessment and did not correlate with any other dimension from either of the other methods. Furthermore, it correlated with only two of the 28 individual behavioral variables. Therefore, there is limited evidence of convergent or ecological validity for this personality dimension. In other Barbary macaque research, Confidence was ecologically validated via its positive correlation with dominance rank, that is, higher ranking individuals were seen as more "confident" (Konečná et al., 2012). In crab eating macaques (*Macaca fasicularis*), age was negatively correlated with trait assessment-derived "Impulsiveness" and "Arousability" (Uher, Werner, et al., 2013). Such results suggest personality dimensions, such as Neuroticism, could be ecologically validated by metrics and individual characteristics rather than just individual behaviors. Developing appropriate methods for this purpose presents a valuable avenue for future ecological validation of personality dimensions. For example, Friendliness and its analogues could be validated through centrality in social networks as proposed by Wilson, Krause, Dingemanse, and Krause (2013).

It is important to note that the forms of behavioral coding and trait assessment used in our study differ methodologically, and the results of our tests of convergent and ecological validity may reflect this. As highlighted already, questionnaires can be built based on species-specific behaviors and thus would be more methodologically similar to behavioral coding approaches (Uher, 2008; Uher, Addessi, et al., 2013; Uher & Asendorpf, 2008; Uher & Visalberghi, 2016; Uher, Werner, et al., 2013). Nevertheless, our aim was to compare popular methods, even if they differed methodologically. Furthermore, as already highlighted, our results show that "top-down" trait assessment and a "bottom-up" behavioral coding can be considered complementary and identify elements of personality a singular approach could not, which is an important consideration for future personality quantification research.

In terms of practicality in a wild setting, data collection using questionnaires is time efficient (assuming there are raters available with sufficient knowledge of the subjects), but may require a large number of subjects to sufficiently power the statistical reduction of the questionnaire items into broader dimensions of personality (Budaev, 2010). Though we applied statistical approaches to calculate the number of components to extract, our study sample size of 27 subjects for a 51-item questionnaire was low and subsequent analyses low in power. As reviewed in Budaev (2010), the adequate ratio of subjects to items in ordination analysis is contentious, with some estimates as high as 10:1. As most of the human research-derived questionnaires used in animal research consist of around 50 items, future studies should ensure they have the sample sizes to sufficiently power these statistical analyses. In our study, for behavioral coding analyses, 18 behavioral variables were analyzed for 27 subjects; such a ratio is still relatively low in power but highlights that fewer subjects may be required to power behavioral coding analyses appropriately when compared with common questionnaire-based approaches.

In our study, experimental assay-derived personality dimensions did not demonstrate interindividual variation, and for one of these personality dimensions, Exploration, there was no evidence of intraindividual consistency. Neither assay-derived personality demonstrated ecological validity. This might be the result of methodical issues due to ethical considerations and permit restrictions, Boldness and Exploration were not quantified using stimuli that have previously been used successfully in other primate species (e.g., simulated predator presence for Boldness or novel food items for Exploration). Instead we used "milder" stimuli of intergroup encounters and novel objects, which in turn may have yielded subtler or more gradual interindividual differences which, though present, could not be identified using our criteria based on likelihood ratio tests and p values. Such inconsistency in stimuli between studies has been critiqued previously (Carter et al., 2013) and limits the scope for phylogenetic and between study comparisons of Boldness or Exploration (Smith & Blumstein, 2008). Designing appropriate experimental assays for particular personality traits is challenging, particularly when definitions for personality dimensions such as Boldness and Exploration remain contentious and, perhaps, species-specific (Carter et al., 2012a, 2012b, 2013). Furthermore, working with wild animals limits the scope for utilizing the broad range of experimental assays that have been employed in captive animal studies (Freeman et al., 2011; Uher, Addessi, et al., 2013; Uher & Visalberghi, 2016). Our failure to find personality dimensions with intraindividual consistency and interindividual variation from experimental assays should not discourage the use of this method in wild animals, but does serve to highlight the practical difficulties of the approach. In addition to the challenge of choosing an appropriate stimulus, working with wild animals makes it challenging to isolate individuals. Thus, there may be a number of uncontrolled social factors affecting individual responses to stimuli, such as the presence of individuals higher in rank (Cronin, Jacobson, Bonnie, & Hopper, 2017). Experimental assays remain a useful and common quantification method for personality, particularly in wild animals where individuals can be isolated or where more invasive methods involving trapping and release are required to measure personality, such as in avian or rodent species (Carere & Maestripieri, 2013). Indeed, in such species, trait assessment and behavioral coding in unconstrained settings are likely to be impractical due to the difficulty of identifying individuals. Furthermore, these experiments elicit interindividual differences in behaviors that, though rare, are potentially significant in terms of fitness, such as responses to predators. Such interindividual differences should be studied further and considered in relation to broader dimensions of personality identified from methods such as behavioral coding in a nonexperimental setting.

Behavioral coding quantifies personality based on the frequencies and rates of species-specific behavioral variables, potentially limiting the scope to standardize methods for interspecies studies (Adams et al., 2015; Freeman et al., 2011). However, although the behavioral variables may be species-specific, for example, the

"triadic embraces" of Barbary macaques (Hodges & Cortes, 2006) or the "genital–genital rubbing" of bonobos (Garai et al., 2016), the personality dimensions derived from behavioral coding can be readily comparable, for example, "Grooming," "Playfulness," and "Introversion" in bonobos (Garai et al., 2016) and "Excitability," "Sociability," and "Tactility" in our study of Barbary macaques. Cross-species comparisons of behavioral coding-derived dimensions can assess the presence or absence of particular dimensions to determine if they are "universal" within taxa (Uher, 2008). For example, it could be explored whether the "Grooming/Tactility" dimension of bonobos/Barbary macaques is found in other primate species, as well as examining the differences in the behaviors that constitute these dimensions. In addition, it could be explored why "tactile" Barbary macaques engage in high rates of self-directed behaviors but "grooming" bonobos do not. Alternatively, future research using behavioral coding could move toward a more standardized framework of behaviors to include in analyses in a similar way to the standardized items appearing in questionnaires. Under such a framework, species-specific behaviors would be included in broader categories, such as "brief affiliation" for the aforementioned triadic embraces and genital–genital rubbing, and derived dimensions would have a comparative power approaching those derived from trait assessments.

Conclusions

Studying animal personality in wild animals offers exciting opportunities to explore how personality has evolved and is maintained in different species in settings most reflective of their evolutionary history. Although our study focuses on only one species in a wild setting, it highlights some of the practical issues of common personality quantification methods, as well as the nonequivalence of their results, arising from important methodological differences between the approaches. Utilizing experimental assays for personality quantification in wild animal research faces various logistical challenges and ethical constraints that limit their use compared with their successful utilization in captive research. Trait assessment has been a popular but contentious approach in personality research, as it is inherently biased to infer intraindividual consistency and interindividual variation where these may not exist. Behavioral coding intrinsically has ecological validity, as it is based on objective, nonmanipulated behavior. It also appears to generate personality dimensions that can be compared across species, particularly within groups of more closely related species, such as primates. However, trait assessment is a relatively simple method to implement and can still complement the quantification of personality of behavioral coding. As already highlighted, our use of both behavioral coding and trait assessment revealed a subtlety to Sociability, which may have been lost by using trait assessment alone, and only trait assessment was able to identify Neuroticism. In addition to highlighting practical issues, our results highlight the methodological differences between the three most popular personality quantification methods within the current literature, suggesting that future comparative work with wild animals should focus on comparisons within methodologies. Finally, our comparison of methods was conducted in a wild, relatively large primate species, in which trait assessment and behavioral coding are feasible due to the ease of identifying specific individuals. Future comparisons of personality quantification methods should also focus on the development of methods that can be implemented in less conspicuous wild animals to expand our knowledge of personality in these species in a wild setting.

References

Adams, M. J., Majolo, B., Ostner, J., Schülke, O., De Marco, A., Thierry, B., . . . Weiss, A. (2015). Personality structure and social style in macaques. *Journal of Personality and Social Psychology, 109,* 338–353. http://dx.doi.org/10.1037/pspp0000041

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour, 49,* 227–266. http://dx.doi.org/10.1163/156853974X00534

Bergvall, U. A., Schäpers, A., Kjellander, P., & Weiss, A. (2011). Personality and foraging decisions in fallow deer, Dama dama. *Animal Behaviour, 81,* 101–112. http://dx.doi.org/10.1016/j.anbehav.2010.09.018

Budaev, S. V. (2010). Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. *Ethology, 116,* 472–480. http://dx.doi.org/10.1111/j.1439-0310.2010.01758.x

Capitanio, J. P. (1999). Personality dimensions in adult male rhesus macaques: Prediction of behaviors across time and situation. *American Journal of Primatology, 47,* 299–320. http://dx.doi.org/10.1002/(SICI)1098-2345(1999)47:4<299::AID-AJP3>3.0.CO;2-P

Carere, C., & Maestripieri, D. (2013). *Animal personalities: behavior, physiology, and evolution*. Chicago, IL: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226922065.001.0001

Carter, A. J., Feeney, W. E., Marshall, H. H., Cowlishaw, G., & Heinsohn, R. (2013). Animal personality: What are behavioural ecologists measuring? *Biological Reviews of the Cambridge Philosophical Society, 88,* 465–475. http://dx.doi.org/10.1111/brv.12007

Carter, A. J., Marshall, H. H., Heinsohn, R., & Cowlishaw, G. (2012a). How not to measure boldness: Novel object and antipredator responses are not the same in wild baboons. *Animal Behaviour, 84,* 603–609. http://dx.doi.org/10.1016/j.anbehav.2012.06.015

Carter, A. J., Marshall, H. H., Heinsohn, R., & Cowlishaw, G. (2012b). Evaluating animal personalities: Do observer assessments and experimental tests measure the same thing? *Behavioral Ecology and Sociobiology, 66,* 153–160. http://dx.doi.org/10.1007/s00265-011-1263-6

Coppens, C. M., de Boer, S. F., & Koolhaas, J. M. (2010). Coping styles and behavioural flexibility: Towards underlying mechanisms. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 365,* 4021–4028. http://dx.doi.org/10.1098/rstb.2010.0217

Corr, P. J., Pickering, A. D., & Gray, J. A. (1995). Personality and reinforcement in associative and instrumental learning. *Personality and Individual Differences, 19,* 47–71. http://dx.doi.org/10.1016/0191-8869(95)00013-V

Cronin, K. A., Jacobson, S. L., Bonnie, K. E., & Hopper, L. M. (2017). Studying primate cognition in a social setting to improve validity and welfare: A literature review highlighting successful approaches. *PeerJ, 5,* e3649. http://dx.doi.org/10.7717/peerj.3649

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., . . . Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39,* 155–164. http://dx.doi.org/10.1111/j.1547-5069.2007.00161.x

Dinno, A. (2012). *paran: Horn's test of principal components/factors*. Retrieved from https://CRAN.R-project.org/package=paran

DiRienzo, N., & Montiglio, P. O. (2015). Four ways in which data-free papers on animal personality fail to be impactful. *Frontiers in Ecology and Evolution, 3,* 23.

Fooden, J. (2007). *Systematic review of the Barbary macaque, Macaca sylvanus (Linnaeus, 1758)*. Chicago, IL: Field Museum of Natural History. http://dx.doi.org/10.5962/bhl.title.14256

Freeman, H. D., Brosnan, S. F., Hopper, L. M., Lambeth, S. P., Schapiro, S. J., & Gosling, S. D. (2013). Developing a comprehensive and com-

parative questionnaire for measuring personality in chimpanzees using a simultaneous top-down/bottom-up design. *American Journal of Primatology, 75,* 1042–1053. http://dx.doi.org/10.1002/ajp.22168

Freeman, H. D., & Gosling, S. D. (2010). Personality in nonhuman primates: A review and evaluation of past research. *American Journal of Primatology, 72,* 653–671. http://dx.doi.org/10.1002/ajp.20833

Freeman, H., Gosling, S. D., & Schapiro, S. J. (2011). Comparison of methods for assessing personality in nonhuman primates. In A. Weiss, J. E. King, & L. Murray (Eds.), *Personality and temperament in nonhuman primates* (pp. 17–40). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-0176-6_2

Garai, C., Weiss, A., Arnaud, C., & Furuichi, T. (2016). Personality in wild bonobos (*Pan paniscus*). *American Journal of Primatology, 78,* 1178–1189. http://dx.doi.org/10.1002/ajp.22573

Gosling, S. D. (2008). Personality in non-human animals. *Social and Personality Psychology Compass, 2,* 985–1001. http://dx.doi.org/10.1111/j.1751-9004.2008.00087.x

Gray, J. A., & McNaughton, N. (2000). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. New York, NY: Oxford University Press.

Hodges, J. K., & Cortes, J. E. (2006). *The Barbary macaque: Biology, management, and conservation*. Nottingham, United Kingdom: Nottingham University Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185. http://dx.doi.org/10.1007/BF02289447

Iwanicki, S., & Lehmann, J. (2015). Behavioral and trait rating assessments of personality in common marmosets (*Callithrix jacchus*). *Journal of Comparative Psychology, 129,* 205–217. http://dx.doi.org/10.1037/a0039318

Jung, S., & Lee, S. (2011). Exploratory factor analysis for small samples. *Behavior Research Methods, 43,* 701–709. http://dx.doi.org/10.3758/s13428-011-0077-9

King, J. E., & Figueredo, A. J. (1997). The five-factor model plus dominance in chimpanzee personality. *Journal of Research in Personality, 31,* 257–271. http://dx.doi.org/10.1006/jrpe.1997.2179

Konečná, M., Weiss, A., Lhota, S., & Wallner, B. (2012). Personality in Barbary macaques (*Macaca sylvanus*): Temporal stability and social rank. *Journal of Research in Personality, 46,* 581–590. http://dx.doi.org/10.1016/j.jrp.2012.06.004

Koolhaas, J. M., Korte, S. M., De Boer, S. F., Van Der Vegt, B. J., Van Reenen, C. G., Hopster, H., . . . Blokhuis, H. J. (1999). Coping styles in animals: Current status in behavior and stress-physiology. *Neuroscience and Biobehavioral Reviews, 23,* 925–935. http://dx.doi.org/10.1016/S0149-7634(99)00026-3

Koski, S. E. (2011). Social personality traits in chimpanzees: Temporal stability and structure of behaviourally assessed personality traits in three captive populations. *Behavioral Ecology and Sociobiology, 65,* 2161–2174. http://dx.doi.org/10.1007/s00265-011-1224-0

Kubinyi, E., Gosling, S. D., & Miklósi, Á. (2015). A comparison of rating and coding behavioural traits in dogs. *Acta Biologica Hungarica, 66,* 27–40. http://dx.doi.org/10.1556/ABiol.66.2015.1.3

Maréchal, L., MacLarnon, A., Majolo, B., & Semple, S. (2016). Primates' behavioural responses to tourists: Evidence for a trade-off between potential risks and benefits. *Scientific Reports, 6,* 32465. http://dx.doi.org/10.1038/srep32465

McDougall, P. T., Réale, D., Sol, D., & Reader, S. M. (2006). Wildlife conservation and animal management: Causes and consequences of evolutionary change for captive reintroduced and wild populations. *Animal Conservation, 9,* 39–48. http://dx.doi.org/10.1111/j.1469-1795.2005.00004.x

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society, 85,* 935–956.

Neumann, C., Agil, M., Widdig, A., & Engelhardt, A. (2013). Personality of wild male crested macaques (*Macaca nigra*). *PLoS ONE, 8,* e69383. http://dx.doi.org/10.1371/journal.pone.0069383

Pederson, A. K., King, J. E., & Landau, V. I. (2005). Chimpanzee (*Pan troglodytes*) personality predicts behaviour. *Journal of Research in Personality, 39,* 534–549. http://dx.doi.org/10.1016/j.jrp.2004.07.002

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2016). nlme: Linear and nonlinear mixed effects models. *R package version, 3,* 1–137. Retrieved from https://CRAN.R-project.org/package=nlme

Réale, D., Reader, S. M., Sol, D., McDougall, P. T., & Dingemanse, N. J. (2007). Integrating animal temperament within ecology and evolution. *Biological Reviews of the Cambridge Philosophical Society, 82,* 291–318. http://dx.doi.org/10.1111/j.1469-185X.2007.00010.x

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych Version=1.8.4

Schino, G., Perretta, G., Taglioni, A. M., Monaco, V., & Troisi, A. (1996). Primate displacement activities as an ethopharmacological model of anxiety. *Anxiety, 2,* 186–191. http://dx.doi.org/10.1002/(SICI)1522-7154(1996)2:4<186::AID-ANXI5>3.0.CO;2-M

Seaman, S. C., Davidson, H. P. B., & Waran, N. K. (2002). How reliable is temperament assessment in the domestic horse (*Equus caballus*)? *Applied Animal Behaviour Science, 78,* 175–191. http://dx.doi.org/10.1016/S0168-1591(02)00095-3

Semple, S., Harrison, C., & Lehmann, J. (2013). Grooming and anxiety in Barbary macaques. *Ethology, 119,* 779–785. http://dx.doi.org/10.1111/eth.12119

Seyfarth, R. M., Silk, J. B., & Cheney, D. L. (2012). Variation in personality and fitness in wild female baboons. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 16980–16985. http://dx.doi.org/10.1073/pnas.1210780109

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Sih, A., Bell, A., & Johnson, J. C. (2004). Behavioral syndromes: An ecological and evolutionary overview. *Trends in Ecology and Evolution, 19,* 372–378. http://dx.doi.org/10.1016/j.tree.2004.04.009

Smith, B. R., & Blumstein, D. T. (2008). Fitness consequences of personality: A meta-analysis. *Behavioral Ecology, 19,* 448–455. http://dx.doi.org/10.1093/beheco/arm144

Sussman, A. F., Ha, J. C., Bentson, K. L., & Crockett, C. M. (2013). Temperament in rhesus, long-tailed, and pigtailed macaques varies by species and sex. *American Journal of Primatology, 75,* 303–313. http://dx.doi.org/10.1002/ajp.22104

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397–423. http://dx.doi.org/10.1006/jrpe.2000.2292

Uher, J. (2008). Comparative personality research: Methodological approaches. *European Journal of Personality, 22,* 427–455. http://dx.doi.org/10.1002/per.680

Uher, J., Addessi, E., & Visalberghi, E. (2013). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (*Cebus apella*). *Journal of Research in Personality, 47,* 427–444. http://dx.doi.org/10.1016/j.jrp.2013.01.013

Uher, J., & Asendorpf, J. B. (2008). Personality assessment in the Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality, 42,* 821–838. http://dx.doi.org/10.1016/j.jrp.2007.10.004

Uher, J., & Visalberghi, E. (2016). Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments.

Journal of Research in Personality, 61, 61–79. http://dx.doi.org/10.1016/j.jrp.2016.02.003

Uher, J., Werner, C. S., & Gosselt, K. (2013). From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods. *Journal of Research in Personality, 47,* 647–667. http://dx.doi.org/10.1016/j.jrp.2013.03.006

Vazire, S., Gosling, S. D., Dickey, A. S., & Schaprio, S. J. (2007). Measuring personality in nonhuman animals. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 190–206). New York, NY: Guilford Press.

Weiss, A. (2017). A human model for primate personality. *Proceedings of the Royal Society B Biological Sciences, 284,* 20171129. http://dx.doi.org/10.1098/rspb.2017.1129

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology, 75,* 1182–1189. http://dx.doi.org/10.1111/j.1365-2656.2006.01141.x

Wilcox, R. R. (1994). The percentage bend correlation coefficient. *Psychometrika, 59,* 601–616. http://dx.doi.org/10.1007/BF02294395

Wilcox, R. R., & Schönbrodt, F. D. (2009). *The WRS package for robust statistics in R (version 0.25.2)*. Retrieved from https://github.com/nicebread/WRS

Wilson, A. D. M., Krause, S., Dingemanse, N. J., & Krause, J. (2013). Network position: A key component in the characterization of social personality types. *Behavioral Ecology and Sociobiology, 67,* 163–173. http://dx.doi.org/10.1007/s00265-012-1428-y

Received November 17, 2017
Revision received October 8, 2018
Accepted October 9, 2018 ■