Current Zoology 60 (2): 162–169, 2014

A critical evaluation of subjective ratings: Unacquainted observers can reliably assess certain personality traits

Matthew B. PETELLE[1*], Daniel T. BLUMSTEIN[1, 2]

[1] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-1606, USA

[2] Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA

Abstract Methods to measure consistent individual differences in behavior (i.e. animal personality) fall into two categories, subjective ratings and behavioral codings. Ratings are seldom used despite being potentially more efficient than codings. One potential limitation for the use of ratings is that it is assumed that long-term observers or experts in the field are required to score individuals. This can be problematic in many cases, especially for long-term ecological studies where there is high turnover in personnel. We tested whether raters who were unacquainted with subjects could produce reliable and valid personality assessments of yellow-bellied marmots *Marmota flaviventris*. Two raters, previously unacquainted with individuals and marmot behavior, scored 130 subjects on fifteen different adjectives in both open-field (OF) and mirror image stimulation (MIS) trials. Eight OF and nine MIS adjectives were reliable as indicated by both a high degree of intra-observer and inter-observer reliability. Additionally, some ratings were externally valid, correlating with behavioral codings. Our data suggest that activity/exploration and sociability can be a reliable and valid measurement of personality traits in studies where raters were unacquainted with subjects. These traits are observable with the personality tests we used; otherwise researchers using unacquainted raters should be cautious in the tests they employ [*Current Zoology* 60 (2): 162–169, 2014].

Keywords Animal personality, Behavior codings, Marmots, Subjective ratings

Animal personality (i.e., consistent individual differences in behavior) has been documented in numerous domestic and wild species (Gosling, 2001; Réale et al., 2007) and from invertebrates to vertebrates (Gosling, 2001; Hensley et al., 2012; Mather and Logue, 2013). Behavioral codings and subjective ratings are two methods used to quantify personality (Gosling, 2001; Vazire et al., 2007). Both methods are applicable for use in the animals' home environments or in behavioral tests, such as open-field and novel object tests.

Behavioral codings measure the presence/absence, frequency, and/or duration of specific postures or behaviors, whereas subjective ratings use observers to score individuals based on a list of adjectives. In studies of non-human animals, codings are more commonly used because of their perceived objectivity and lack of human bias. Whereas ratings are used to a lesser extent, they are seen as a more holistic way to assess personality; additionally they are seen as more efficient than behavioral codings because of how quickly they can be conducted once an observer is trained (Vazire et al., 2007).

Despite the potential advantages of ratings, short- and long-term studies with high personnel turnover may not utilize this method because of the notion that raters must be well acquainted with subjects in order to accurately assess personality. Consequently, in the majority of studies that use ratings, observers are commonly breeders, trainers, or long-term animal care providers (Carter et al., 2012; Fratkin et al., 2013; Uher and Asendorpf, 2008; Wilsson and Sinn, 2012). This can be problematic for long-term ecological studies where there is high turnover in personnel. Additionally, a potential consequence of using well-acquainted observers is the potential for confirmation bias due to preconceptions that raters may have of animal subjects (Highfill et al., 2010). Surprisingly, we do not yet fully understand how acquaintance with subjects may influence ratings in either captive or wild studies.

There is research to suggest that while reliability of measures increases with level of acquaintance, raters less acquainted with subjects can also score subjects satisfactorily (Martau et al., 1985; Wemelsfelder et al., 2000). In Martau et al.'s (1985) study of 12 Japanese macaques *Macaca fuscata*, well acquainted and less acquainted raters scored individuals. Less acquainted

Received May 15, 2013; accepted Aug. 3, 2013.

* Corresponding author. E-mail: mpetelle@ucla.edu

raters observed subjects for up to 1 hour a day for 5 days before rating those same individuals while familiar raters observed subjects for 2 hour a day for up to a month. Although well-acquainted raters had higher inter-observer agreements, raters less familiar with the animals were still able to achieve high inter-observer agreement. However, in this case, the less acquainted raters had a level acquaintance with test subjects typically not achievable in many field studies. Wemelsfelder et al. (2000) found that multiple unacquainted observers had clear agreement in how they qualitatively described pig behavior, but these observer ratings were not tested for validity.

Before a measurement can be informative, it must be both reliable and valid. Reliability can be assessed with two methods: inter-rater agreement, and test-retest reliability (Vazire et al., 2007). Inter-rater agreement, typically measured by intra-class correlation coefficients, is an index of how well multiple observers agree in their personality ratings of an individual. Gosling (2001), in an extensive review of animal personality, found that inter-observer agreement in animals was comparable to reliability estimates in the human personality literature (grand mean 0.52). Furthermore, reliability is also assessed through test-retest reliability, or repeatability. This statistic describes how consistent an individual's personality score is across time. Repeatability depends upon taxa, sex, age, laboratory vs. field, and length between tests (Bell et al., 2009; Gosling, 2001). Gosling (2001) found that test-retest reliabilities were generally high with a range from 0.31–0.90.

Validity is an index of how well a measurement is describing what it is supposed to measure (Vazire et al., 2007). Validity can be assessed with a number of techniques. One common method to assess the external validity of ratings is to compare them to behavioral codings that are associated with that particular adjective (Gosling, 2001). For example, an individual rated as being highly sociable may spend more time at a mirror during a mirror image stimulation test or be more embedded in a social network. There are several examples of acquainted raters, up to two hours pre-trial observation, assigning subjective scores that externally predict an individuals behavioral coding in other tests (Fox and Millam, 2010; Barnard et al., 2012; Carter et al., 2012).

Here we test the reliability and validity of ratings on a long-term study of yellow-bellied marmots *Marmota flaviventris* with raters that were unacquainted with individuals and, before training, with their species-specific behavior. If subjective ratings are reliable and valid, personnel unacquainted with subjects can use them in standardized test situations.

1 Materials and Methods

1.1 Study area and system

We conducted experiments in and around the Rocky Mountain Biological Laboratory (RMBL, 38°57'N, 106°59'W), Gothic, CO, USA in 2010 (May–August). Marmots were regularly live-trapped and transferred to a cloth, handling bag where sex, reproductive status, and mass were determined. Marmots were marked with permanent ear tags for identification as well as unique fur marks (with Nyanzol fur dye) for observation from afar. Almost all marmots from the population are trapped at least once during the active season (mid-April to mid-September). Yellow-bellied marmots from this population have been previously shown to have personalities (Armitage and Van Vuren, 2003; Blumstein, et al., 2012; Svendsen and Armitage, 1973).

1.2 General tests

Open-field (OF) and mirror image stimulation (MIS) trials were conducted in an arena measuring 91.4 cm^3 made of 0.47 cm opaque PVC sheeting with a wire mesh top to prevent escape. A mirror (30.5×61.0 cm) was placed at the base of one side of the arena and covered with an opaque sliding door. A door (61.0 cm^2) was cut out of the opposite side. Sixteen (22.9 cm^2) squares were drawn on the bottom of the arena in a grid to record location of individuals. The arena was placed under a canopy for shade and to standardize the light environment. Trials were video-recorded (Sharp Mini DV Digital Camera) from above, for later scoring. We gently released individuals from the top of the arena. The first three minutes were considered the open-field test. During the open field test, individuals were allowed to freely explore the arena. The OF setup is similar to one used for testing personality in Alpine marmots *Marmota marmota* (Constantini et al. 2012; Ferrari et al., 2013; Réale personal communication). Immediately after the first three minutes, the MIS trial began by removing the sliding door to expose the mirror. Upon trial completion, we placed a Tomahawk trap within the door and urged the marmot inside. We returned individuals to the location originally trapped and cleaned the arena with a vinegar and water solution before the next trial. In total, we performed 205 open field (OF) and mirror image stimulation (MIS) trials on 130 individuals (32 juvenile females, 30 juvenile males, 20 yearling females, 22 yearling males, 16 adult females, and 10 adult males). Seventy-seven animals were tested twice, with six of those animals being tested a third time. All trials were included in analyses. Marmots were trapped opportunistically, and therefore individuals were tested sporadically throughout the active season. We used open field and mirror image stimulations because they are standardized; rating individual personality in natural settings would require raters to understand both the social and environmental context in which the behavior was recorded. Furthermore, we included juvenile individuals because they have been shown to exhibit personality (Armitage, 1986a; unpublished data).

1.3 Personality measurements

1.3.1 Subjective ratings

Videos of trials were sorted and viewed by sex and age categories to control for sex-specific ontogenetic variation in behavior. Thus, all scores were relative to the same age/sex category and not between all individuals. We chose 15 adjectives (Table 1), some from a previous list used on rhesus macaques (Capitanio, 1999), and others that have been used recently on studies of heteromyid rodents with high intraclass correlations (L. Baker, pers. comm., University of British Columbia). Marmots were scored on a scale from 1–7 in increments of 0.25, where 1 describes the individual as not exhibiting the trait, while 7 describes the trait being fully exhibited. This is similar to the method employed by Capitanio (1999), except we allowed for a finer division of ratings.

Table 1 Intra-class correlation coefficients of adjectives used to describe yellow-bellied marmots in open-field and mirror image stimulation tests

Adjective	OF		MIS	
	ICC	*P*	ICC	*P*
Active	**0.577**	**<0.0001**	**0.673**	**<0.0001**
Aggressive	-0.001	0.503	**0.511**	**<0.0001**
Apprehensive	**0.231**	**0.031**	-0.199	0.902
Cautious	0.013	0.463	-0.039	0.606
Confident	0.003	0.493	**0.271**	**0.012**
Curious	**0.577**	**<0.0001**	**0.567**	**<0.0001**
Excitable	**0.51**	**<0.0001**	**0.426**	**<0.0001**
Fearful	0.165	0.099	0.025	0.428
Irritable	0.109	0.205	0.153	0.118
Oppositional	**0.635**	**<0.0001**	**0.366**	**0.001**
Playful	0.063	0.322	**0.582**	**<0.0001**
Protective	**0.741**	**<0.0001**	**0.791**	**<0.0001**
Deliberate	0.181	0.078	0.173	0.088
Solitary	**0.697**	**<0.0001**	**0.539**	**<0.0001**
Strong	**0.264**	**0.015**	0.067	0.311

Significant values in bold.

Two raters (UCLA undergraduates) were chosen from a pool of undergraduate applicants. Neither rater had observed marmot behavior prior to watching these trials. Both raters were given the adjectives and viewed trials from juvenile, female marmots. After viewing, raters and MP discussed the adjectives and the behaviors that potentially constituted each adjective. Each rater scored 15 randomly selected juvenile female OF/MIS trials and scored them up to five times until they had high intra-rater agreement. High intra-rater agreement was defined as scores having a $r_S > 0.90$. Raters watched, but did not score, 10~15 trials of the subsequent sex/age category (e.g., juvenile females; juvenile males; yearling females, etc.) to understand differences in behavior between individuals and the previous category. All trials were watched and rated on computers at UCLA.

1.3.2 Quantitative codings

Behavior was scored using the event recorder JWatcher (Blumstein and Daniel, 2007) to calculate the number of events and the proportion of time spent walking, looking (quadrapedal and bipedal), jumping, alarm calling, smelling or sniffing, and, for MIS only, scratching, pawing, or pressing their nose against the mirror. Additionally, activity was scored by counting the number of lines crossed using the nose of the subject as an indicator of its location, proportion of squares visited, and for MIS only, the proportion of time spent in front of the mirror and on the mirrored half of the arena (Table 1). Prior to scoring trials, scorers were trained to have high intra- and inter-observer agreement ($r > 0.95$). To ensure high intra- and inter-observer agreement in quantifying behavior, MP scored a trial multiple times until the frequencies of all behaviors were equal and total durations of behaviors were within 5% between each scoring events. This method was carried out for five trials. Other scorers had to record the same behavioral frequency and estimated durations to ensure inter-observer agreement. Raters did not code behaviors. This was done by MP and other trained UCLA undergraduates.

1.4 Analyses

1.4.1 Inter-rater and test-retest reliability

All individual marmots were grouped for analysis. We analyzed OF and MIS separately. To assess inter-rater reliability for each of the 15 adjectives, we used an intraclass correlation coefficient (ICC) using a two way mixed model that measured consistency because both coders rated all individuals (Shrout and Fleis, 1979). Adjectives that had a significant ICC ($P < 0.05$) were included in future analyses. All further analyses were based on a single rating that was obtained by averaging rater scores.

We assessed test-retest reliability using individual repeatability. To obtain repeatability for individual marmots, we fit a linear mixed effects model for each adjective with age category, sex, age category * sex, rater, and trial as fixed effects, and individual as a random effect. Age category and sex have been found to influence other behaviors, including personality dimensions (Blumstein et al., 2012). We included the scores from both raters in the model and included rater as a fixed effect. We also included trial to control for habituation effects. We estimated the repeatability of each adjective by dividing the variance explained by the individual by the total phenotypic variance explained by the model. Significance of repeatability was estimated with a log-likelihood ratio test (Pinheiro and Bates, 2000). Only adjectives that had significant inter-rater and test-retest reliability were included in rating validity.

1.4.2 Validity of ratings

We tested rating external validity by including all ratings with behavioral codings in a principal component analysis. Ratings and codings that are correlated load onto the same component (J.G.A. Martin, pers. comm., University of Aberdeen). We used a Varimax rotation to aid in interpretation. For component selection, we conducted a parallel analysis with 1,000 randomly selected data sets with 95% confidence intervals for both OF and MIS PCAs. Significant components were kept for further interpretation (O'Connor, 2000). Variables with values > |0.40| were used to interpret factors. All analyses were conducted in SPSS v. 18.0 (Chicago, Il) and R 2.14.0 (R Development Core Team 2011) with the package lme4 (Bates et al., 2011). We set our alpha to 0.05.

2 Results

2.1 Inter-rater reliability and test-retest reliability

Eight of the fifteen adjectives for OF had significant ICCs. Additionally, nine of the fifteen adjectives for MIS had significant ICCs (Table 1). Six of the eight OF adjectives had significant repeatability: active (r = 0.182, LRT = 19.853, P < 0.0001), curious (r = 0.123, LRT = 7.838, P = 0.005), excitable (r = 0.170, LRT = 11.482, P = 0.0007), oppositional (r = 0.321, LRT = 40.132, P < 0.0001), protective (r = 0.369, LRT = 55.315, P < 0.0001), and solitary (r = 0.221, LRT = 23.715, P < 0.0001). All nine MIS adjectives had significant repeatability: active (r = 0.330, LRT = 38.213, P < 0.0001), aggressive (r = 0.185, LRT = 14.258; P = 0.0002), confident (r = 0.111, LRT = 5.318, P = 0.021), curious (r = 0.220, LRT = 20.872, P < 0.0001), excitable (r = 0.258, LRT = 28.894, P < 0.0001), oppositional (r = 0.156, LRT = 10.317, P = 0.001); playful (r = 0.408, LRT = 41.078, P < 0.0001), protective (r = 0.472, LRT = 75.956, P < 0.0001), and solitary (r = 0.222, LRT = 19.258, P < 0.0001).

2.2 Validity of ratings

Principle component analysis for the open field test extracted two components explaining 57.13% of the variation. The first component was interpreted as an activity and exploration factor. It was loaded with the proportion of boxes visited, number of lines crossed, number of jumps, number of rear looks, number of walks, proportion of time looking, proportion of time in rear look, proportion walking, active, curious, oppositional, protective, and solitary. The second component was also interpreted as an exploration factor with number of sniffs and proportion of time sniffing as significant variables (Table 2).

Principle component analysis for the mirror image stimulation test extracted five components explaining 69.19% of the variation. The first component was interpreted as an activity and exploration factor. It was

Table 2 Summary of principle component analysis for open-field (OF)

OF	Component	
Behaviors/adjectives	Activity/Exploration	Exploration
Active	**0.619**	0.246
Curious	**0.43**	0.285
Excitable	0.08	0.025
Oppositional	**0.575**	-0.004
Protective	**-0.782**	-0.123
Solitary	**-0.769**	-0.126
Prop boxes visited	**0.574**	0.363
N lines crossed	**0.764**	0.325
N alarm calls	-0.061	0.157
N jumps	**0.592**	-0.265
N looks	-0.117	0.387
N sniff/smell	0.254	**0.912**
N rear looks	**0.861**	-0.248
N walks	**0.795**	0.254
Prop look	**-0.905**	-0.326
Prop sniff/smell	0.135	**0.917**
Prop rear look	**0.831**	0.017
Prop walk	**0.845**	0.182

Variables with coefficients larger than |0.4| are highlighted in bold.

loaded with proportion of boxes visited, number of lines, number of looks, number of sniffs, number of walks, proportion of time looking, sniffing, and walking. Also included are active and curious adjectives. The second component was interpreted as a sociability component with proportion of time spent at the mirror, proportion of time spent on the mirrored half, number of scratches or nose touches, proportion of time scratching or nose touching, active and curious. The third component was also interpreted as a sociability component with active, aggressive, confident, oppositional, playful, protective, and solitary adjectives. The fourth component was also associated with exploration. It was loaded with number of rear looks, proportion of time looking and proportion of time rear looking. The fifth component was labeled as an excitability component with number of jumps and active, excitable, and oppositional adjectives (Table 3).

3 Discussion

Numerous studies have already found that acquainted raters can assess personality (Gosling, 2001), thus, this study investigates whether unacquainted raters can reliably and validly score personality traits. We found that subjective ratings by unacquainted raters were reliable and valid for two personality traits--activity/exploration and sociability. Specifically, subjective ratings within open-field tests were used to identify an activity/exploration personality trait while mirror image stimulation identified both an activity and a sociability personality trait. These results suggest that in certain standardized tests, subjective ratings made by people not intimately familiar with the subjects can be a useful method to quantify personality dimensions.

3.1 Reliability of personality measurements

The majority of our adjectives had significant inter-rater reliabilities. Six adjectives with significant ICCs were shared across both OF and MIS tests. This suggests that these adjectives are perhaps easier to recognize within and across situations. Active, curious, excitable, protective, and solitary were all found to have

Table 3 Summary of principle component analysis for mirror image stimulation

MIS	Component				
Behaviors/ adjectives	Activity/ Exploration	Sociability	Sociability 2	Exploration 2	Excitability
Active	0.554	0.214	0.536	0.128	0.276
Aggressive	0.125	0.1	0.52	0.121	0.689
Confident	0.221	0.201	0.756	-0.059	0.199
Curious	0.472	0.413	0.281	0.012	0.076
Excitable	0.048	0.162	-0.126	-0.05	0.753
Oppositional	0.133	-0.108	0.438	0.089	0.68
Playful	0.211	0.346	0.671	0.011	0.023
Protective	-0.375	-0.285	-0.634	-0.336	-0.127
Solitary	-0.24	-0.181	-0.604	-0.385	0.043
Prop boxes visited	0.776	0.214	0.079	0.252	0.223
N lines crossed	0.693	0.259	0.101	0.199	0.19
Prop at mirror	0.058	0.847	0.21	0.024	0.082
Prop mirror half	0.101	0.684	0.191	0.141	0.028
N alarm calls	-0.115	0.001	-0.041	-0.051	0.017
N jumps	0.075	0.18	0.078	0.381	0.652
N looks	0.48	0.246	0.103	-0.025	0.073
N sniff/smell	0.846	0.057	0.241	0.166	-0.021
N scratch/paw	0.396	0.73	0.226	0.042	0.158
N rear looks	0.298	-0.042	0.078	0.89	0.113
N walks	0.748	0.282	0.217	0.257	0.125
Prop looks	-0.516	-0.339	-0.181	-0.647	-0.118
Prop sniff/smell	0.824	0.043	0.222	0.121	-0.061
Prop scratch/smell	0.237	0.797	0.158	0.019	0.135
Prop rear look	0.234	0.019	0.063	0.883	0.118
Prop walk	0.741	0.08	0.266	0.451	0.054

Variables with coefficients larger than |0.4| are highlighted in bold.

similar, if not higher, inter-rater reliability than other studies (0.62, 0.47, 0.38, 0.38, and 0.43 respectively) (see Gosling, 2001). Other adjectives, however, may not be appropriate for all contexts, thus accounting for differences in reliability scores, or differences between individuals may be too subtle for specific tests or observers to identify (Meagher, 2009). Interestingly, only two of the five adjectives, active and curious, were observed to have high observability across species. Observability refers to how visible a trait is within a given situation or context. We are not sure why the other three adjectives were so high compared to Gosling's (2001) findings, perhaps these adjectives are appropriate for this species within this context, and are thus more observable. Open-field tests, for example, were designed to assess fear and activity, thus it is not surprising that adjectives describing these traits may be easier to rate in this situation. Conversely, both aggression and playfulness are commonly thought of as social attributes and might therefore be more observable in the mirror-image stimulation.

Of those adjectives with significant inter-rater reliabilities we found many of these to be repeatable. Personality, by definition, must be repeatable, and therefore the test-retest reliability is essential to include in any analysis of ratings. As our study shows, adjectives that have high inter-rater agreement are not necessarily repeatable, and thus should not necessarily be viewed as personality traits without further justification. Additionally, our repeatability estimates are generally moderate, but fall within the range of repeatable behaviors (Bell et al., 2009).

We should note that we did not test all individuals multiple times. While this could affect repeatability estimates for linear mixed effects models, Martin et al. (2011) advised that large data sets ($n > 200$) are sufficient to estimate individual differences, and that including individuals with one observation actually increases the power to detect these differences. Therefore, we are confident that our results accurately reflect the test-retest reliability of these adjectives.

We recognize that the use of two raters can result in an overestimate of ICC scores, and therefore our results indicate the upper-limit for reliability in these scores. However, our results suggest that just two raters can reliably score certain adjectives. Studies that use acquainted raters typically rely on one to five raters (Martau et al., 1985; Highfill et al., 2010; Barnard et al., 2012). Moreover, this experiment is part of an ongoing ecological study where high personnel turnover is common. Consequently we have a vested interest in determining if a minimum number of unacquainted raters will suffice in judging personality.

3.2 Validity of subjective ratings

Principle Component Analysis revealed that the five reliable adjectives in the OF test were correlated with behaviors that can often be used to define an activity or exploration trait. Thus, our study suggests that raters, unacquainted with subjects, were able, with minimal training, to use adjectives that describe an active/exploration personality trait during OF tests. Our results are consistent with other studies on Alpine marmots where the first component reveals an activity/exploration trait with movement and upright posture being correlated (Ferrari et al., 2013).

We also found that raters were able to describe activity/exploration within the MIS test along with a sociability component. MIS tests are widely used to assess how individuals interact with an unknown conspecific, and therefore they are often used as a metric of sociability (Armitage, 1986a; Armitage, 1986b). Additionally, we found an excitability component with aggressive, excitable, and oppositional loading significantly with number of jumps. This component was not seen in the OF test, suggesting that these correlated behaviors are related to being exposed to a mirror. Excitability has been shown in a number of studies that use ratings and is common in laboratory studies of rats (Cerbone, 1993; Gosling and John, 1998).

Interestingly, the fact that curious loads positively on two components, activity/exploration and sociability, suggests that subjective ratings provide a broader qualitative description, or holistic view of individuals, which may cover multiple traits (Uher and Asendorpf, 2008). Surprisingly, we found that adjectives that describe sociability-playful and aggressive-were not associated with time spent at the mirror. This suggests that although adjectives such as playful and aggressive can be reliably scored, they are not externally valid in this context to explore sociability.

Adjectives that were not reliably scored, or were reliable and not valid, may result from the tests not being ecologically relevant. These adjectives may be more observable (reliable and valid) if underlying tests are able to expose those underlying traits. Another potential method to pinpoint more relevant adjectives is to have them chosen to reflect traits known to exist in the test species (e.g. Armitage, 1986b and Blumstein et al., 2006). For example, mirror image stimulation codings have previously been used to determine sociability in marmots. These MIS scores were ecologically relevant, correlating with social interactions and reproductive success (Armitage and Van Vuren, 2003). Although these adjectives are useful in describing personality traits in this specific population of marmots, each species and population has different traits and correlation between those traits (Bell and Stamps, 2004; Dingemanse et al., 2007). Thus, a different set of adjectives may be a better indicator of personality traits. Taking a bottom-up approach, or watching individuals in ecologically relevant situations and then listing potential adjectives might be a more effective way of using adjectives (Uher and Asendorpf, 2008). Thus, for long-term ecological studies, personnel well acquainted with the species and individuals in the population should determine adjectives and tests used to define personality traits (Meagher, 2009). This method can potentially be used for a number of taxa including some invertebrates given that the personality traits are highly observable in a standardized test. For example, it may be very easy for unacquainted observers to rate individuals on an activity/exploration axis in an open field test.

Our study suggests that projects with high personnel turnover should be able to effectively use ratings to reduce time and resources to score behaviors and quantify some personality traits provided that raters are properly trained beforehand and subjects are tested in a standardized manner. Those traits studied, however, should be restricted to ones that are explicitly observable. For example, our study shows that OF and MIS tests can be used to identify active and active/sociable traits, but not other traits. Indeed, the reliability of difficult to score traits should be generally scrutinized when relying on expert raters.

Acknowledgments M.B.P. was supported by a GAANN fellowship, a GK-12 fellowship, and the UCLA Department of Ecology and Evolutionary Biology. D.T.B. is supported by the NSF. We thank Julien Martin for statistical help, Liv Baker for additional assistance, and Lynn Fairbanks and Peter Nonacs for comments on previous drafts. We also thank Josh Ryan and Laurel Brewer for rating the individuals, and all the marmoteers, especially Leon Chong, who helped collect data. We also thank three anonymous reviewers for their comments that were helpful in the revision of this manuscript.

References

Armitage KB, 1986a. Individual differences in the behavior of juvenile yellow-bellied marmots. Behav. Ecol. Sociobiol. 18: 419–424.

Armitage KB, 1986b. Individuality, social behavior, and reproductive success in yellow-bellied marmots. Ecology 67: 1186–1193.

Armitage KB, Van Vuren DH, 2003. Individual differences and reproductive success in yellow-bellied marmots. Ethol. Ecol. Evol. 15: 207–233.

Barnard S, Siracusa C, Reisner I, Valsecchi P, Serpell JA, 2012. Validity of model devices used to assess canine temperament in behavioral tests. App. Anim. Behav. Sci. 138: 79–87.

Bates D, Maechler M, Bolker B, 2011. lme4: Linear mixed–effects models using S4 classes. R package version 0.999375-42. http://CRAN.R-project.org/package=lme4

Bell AM, Hankison SJ, Laskowski KL, 2009. The repeatability of behaviour: A meta-analysis. Anim. Behav. 77: 771–783.

Bell AM, Stamps JA, 2004. Development of behavioural differences between individuals and populations of sticklebacks *Gasterosteus aculeatus*. Anim. Behav. 68: 1339–1348.

Blumstein DT, Daniel JC, 2007. Quantifying behavior the JWatcher way. Sunderland: Sinauer Associates Inc.

Blumstein DT, Holland BD, Daniel JC, 2006. Predator discrimination and 'personality' in captive Vancouver Island marmots *Marmota vancouverensis*. Anim. Conserv. 9: 274–282.

Blumstein DT, Petelle MP, Wey TW, 2012. Defensive and social aggression: Repeatable but independent. Behav. Ecol. 24: 457–461.

Capitanio JP, 1999. Personality dimensions in adult male rhesus macaques: Prediction of behaviors across time and situation. Am. J. Primatol. 47: 299–320.

Carter A, Marshall H, Heinsohn R, Cowlishaw G, 2012. Evaluating animal personalities: Do observer assessments and experimental tests measure the same thing? Behav. Ecol. Sociobiol. 66: 153–160.

Cerbone A, Pellicano P, Sadile AG, 1993. Evidence for and against the naples high- and low-excitability rats as genetic model to study hippocampal functions. Neurosci. Biobehav. Rev. 17: 295–303.

Constantini D, Ferrari C, Pasquaretta C, Cavallone E, Carere C et al, 2012. Interplay between plasma oxidative status, cortisol and coping styles in wild alpine marmots *Marmota marmota*. J. Exper. Biol. 215: 374–383.

Dingemanse NJ, Wright J, Kazem AJ, Thomas DK, Hickling R et al., 2007. Behavioural syndromes differ predictably between 12 populations of three-spined stickleback. J. Anim. Ecol. 76: 1128–1138.

Ferrari C, Pasquaretta C, Carere C, Cavallone E, von Hardenberg A et al., 2013. Testing for the presence of coping styles in a wild mammal. Anim. Behav. 85: 1385–1396.

Fox RA, Millam JR, 2010. The use of ratings and direct behavioural observation to measure temperatment traits in cockatiels *Nymphicus hollandicus*. Ethology 116: 59–71.

Fratkin JL, Sinn DL, Patall EA, Gosling SD, 2013. Personality consistency in dogs: A meta-analysis. PLOS One 8(1): doi: 10.1371/journal.pone.0054907

Gosling SD, John OP, 1998. Personality dimensions in nonhuman animals: A cross-species review. Curr. Dir. Psychol. Sci. 8: 69–75.

Gosling SD, 2001. From mice to men: What can we learn about personality from animal research? Psychol. Bull. 127: 45–86.

Highfill F, Hanbury D, Kristiansen R, Kuczaj S, Watson S, 2010.

Rating *vs*. coding in animal personality research. Zoo Biol. 29: 509–516.

Hensley NM, Cook TC, Lang M, Petelle MB, Blumstein DT, 2012. Personality and habitat segregation in giant sea anemones *Condylactis gigantea*. J. Exp. Marine Biol. Ecol. 1–4: 426–427.

Mather JA, Logue D, 2013. The bold and the spineless: Invertebrate personalities. In: Carere C, Maestripieri D ed. Animal Personalities: Behavior, Physiology, and Evolution. Chicago: The University of Chicago Press, 23–53.

Martau PA, Caine NG, Candland DK, 1985. Reliability of the emotions profile index, primate form, with *Papio hamadryas*, *Macaca fuscata*, and two *Saimiri* species. Primates 26: 501–505.

Martin JGA, Nussey DH, Wilson AJ, Réale D, 2011. Measuring individual differences in reaction norms in field and experimental studies: A power analysis of random regression models. Methods Ecol. Evol. 2: 362–374.

Meagher RK, 2009. Observer ratings: Validity and value as a tool for animal welfare research. Appl. Anim. Behav. Sci. 119: 1–14.

Murphy LB, 1978. The practical problems of recognizing and measuring fear and exploration behaviour in the domestic fowl. Anim. Behav. 26 (Part 2): 422–431.

O'Connor BP, 2000. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. Behav. Res. Meth. Instrum. and Comp. 32: 396–402.

Pinheiro JC, Bates DM, 2000. Mixed-effects Models in S and S-PLUS. New York: Springer Verlag.

Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ, 2007. Integrating animal temperament within ecology and evolution. Biol. Rev. 82: 291–318.

Shrout PE, Fleis JL, 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. 86: 420–428.

Svendsen GE, Armitage KB, 1973. Mirror-image stimulation applied to field behavioral studies. Ecology 54: 623–627.

Uher J, Asendorpf JB, 2008. Personality assessment in the great apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. J. Res. Pers. 42: 821–838.

Vazire S, Gosling SD, Dickey AS, Schapiro SJ, 2007. Measuring personality in nonhuman animals. In: Robins RW, Fraley CR, Krueger RF ed. Handbook of Research Methods in Personality Psychology. New York: The Guilford Press, 190–206.

Wemelsfelder F, Hunter EA, Mendl MT, Lawrence AB, 2000. The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. Appl. Anim. Behav. Sci. 67: 193–215.

Wilsson E, Sinn DL, 2012. Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program. Appl. Anim. Behav. Sci. 141: 158–172.