Modelling personality, plasticity and predictability in shelter dogs

Conor Goold*[1] and Ruth C. Newberry[1]

[1]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences

June 2, 2017

Abstract

Behavioural assessments of shelter dogs (*Canis lupus familiaris*) typically comprise standardised test batteries conducted at one time point but test batteries have shown inconsistent predictive validity. Longitudinal behavioural assessments offer an alternative. We modelled longitudinal observational data on shelter dog behaviour using the framework of behavioural reaction norms, partitioning variance into personality (i.e. inter-individual differences in behaviour), plasticity (i.e. individual differences in behavioural change) and predictability (i.e. individual differences in residual intra-individual variation). We analysed data on 3,263 dogs' interactions (N = 19,281) with unfamiliar people during their first month after arrival at the shelter. Accounting for personality, plasticity (linear and quadratic trends) and predictability improved the predictive accuracy of the analyses compared to models quantifying personality and/or plasticity only. While dogs were, on average, highly sociable with unfamiliar people and sociability increased over days since arrival, group averages were unrepresentative of all dogs and predictions made at the individual level entailed considerable uncertainty. Effects of demographic variables (e.g. age) on personality, plasticity and predictability were observed. Behavioural repeatability increased with days since arrival. Our results highlight the value of longitudinal assessments on shelter dogs and identify measures that could improve the predictive validity of behavioural assessments in shelters.

Keywords— inter- and intra-individual differences, behavioural reaction norms, behavioural repeatability, longitudinal behavioural assessment, human-animal interactions.

*Corresponding author: conor.goold@nmbu.no

1 Introduction

Personality, defined by inter-individual differences in average behaviour, represents just one component of behavioural variation of interest in animal behaviour research. Personality frequently describes less than 50% of behavioural variation in animal personality studies [1, 2], leading to the combined analysis of personality with *plasticity*, individual differences in behavioural change [3], and *predictability*, individual differences in residual intra-individual variability [4–8]. Understanding these different sources of behavioural variation simultaneously can be achieved using the general framework of behavioural reaction norms [3, 5], which provides insight into how animals react to fluctuating environments through time and across contexts. More generally, these developments reflect increasing interest across biology in expanding the 'trait space' of phenotypic evolution [9] beyond mean trait differences and systematic plasticity across environmental gradients to include residual trait variation (e.g. developmental instability: [10, 11]; stochastic variation in gene expression: [12]).

Modest repeatability of behaviour has been documented in domestic dogs (*Canis lupus familiaris*), providing evidence for personality variation. For instance, using meta-analysis, Fratkin *et al.* [13] found an average Pearson's correlation of behaviour through time of 0.43, explaining 19% of the behavioural variance between successive time points. However, the goal of personality assessments in dogs is often to predict an individual dog's future behaviour (e.g. working dogs: [14, 15]; pet dogs: [16]) and, thus, it is important not to confuse the stability of an individual's behaviour relative to the behaviour of others with stability of intra-individual behaviour. That is, individuals could vary their behaviour in meaningful ways while maintaining differences from other individuals. As illustrated in Figure 1, a correlation of 0.4 in behaviour across repeated measurements does not preclude individual heterogeneity in plasticity or predictability. When time-related change in dog behaviour has been taken into account, behavioural change at the group-level has been of primary focus (e.g. [16–18]) and no studies have explored the heterogeneity of residual variance within each dog. The predominant focus on inter-individual differences and group-level patterns of behavioural change risks obscuring important individual-level heterogeneity and may partly explain why a number of dog personality assessment tools have been unreliable in predicting future behaviour [14–16, 19].

Of particular concern is the low predictive value of shelter dog assessments for predicting behaviour post-adoption [20–24], resulting in calls for longitudinal, observational models of assessment [24]. Animal shelters are dynamic environments and, for most dogs, instigate an immediate threat to homeostasis as evidenced by heightened hypothalamic-pituitary-adrenal axis activity and an increase in stress-related behaviours (e.g. [25–28]). Over time, physiological and behavioural responses are amenable to change [17, 27, 29]. Therefore, dogs in shelters may exhibit substantial heterogeneity in intra-individual behaviour captured neither by standardised behavioural assessments conducted at one time point [24] nor by group-level patterns of behavioural change. An additional complication is that the behaviour in shelters may not be representative of behaviour outside of shelters. For example, Patronek and Bradley [29] suggested that up to 50% of instances of aggression expressed while at a shelter are likely to be false positives. Such false positives may be captured in estimates of predictability, with individuals departing more from their representative behaviour having

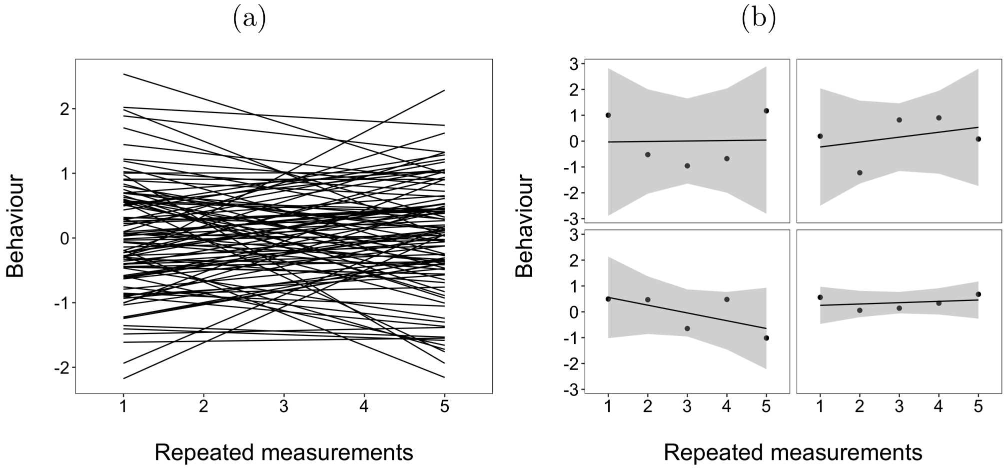


Figure 1: (a) *Reaction norms for 100 simulated individuals measured on five occasions, with a correlation of 0.4 between successive time points.* (b) *Reaction norms and raw data (black points) for four randomly selected individuals; shaded areas represent the residual intra-individual variability or predictability around reaction norm estimates.*

higher residual intra-individual variability (lower predictability) than others. Overall, absolute values of behaviour, such as mean trait values across time (i.e. personality), may account for just part of the important behavioural variation needed to understand and predict shelter dog behaviour. While observational models of assessment have been encouraged, methods to systematically analyse longitudinal data collected at shelters into meaningful formats are lacking.

In this paper, we demonstrate how the framework of behavioural reaction norms can quantify inter- and intra-individual differences in shelter dog behaviour. To do so, we use data on dogs' interactions with unfamiliar people from a longitudinal and observational shelter assessment. As a core feature of personality assessments, how shelter dogs interact with unknown people is of great importance. At one extreme, if dogs bite or attempt to bite unfamiliar people, they are at risk of euthanasia [29]. At the other extreme, even subtle differences in how dogs interact with potential adopters can influence adoption success [30]. Importantly, neither may all dogs react to unfamiliar people in the same way through time at the shelter nor may all dogs show the same day-to-day fluctuation of behaviour around their average behavioural trajectories. These considerations can be examined with behavioural reaction norms.

The analysis of behavioural reaction norms is dependent on the use of hierarchical statistical models for partitioning variance among individuals [3, 5, 6]. Given that ordinal data are common in behavioural research, here, we illustrate how similar hierarchical models can be applied to ordinal data using a Bayesian framework (see also [31]). Apart from distinguishing inter- from intra-individual variation, we place particular emphasis on two desirable properties of the hierarchical modelling approach taken here. First, the property of *hierarchical shrinkage* [32] offers an efficacious way of making inferences about individual-level behaviour when data are highly unbalanced and potentially unrepresentative of a dog's typical behaviour. When data are sparse for certain individuals, hierarchical shrinkage will attenuate their estimates to the group-level estimates. Similarly, if data are unrepresentative of group-level patterns, estimates will be more informed by group-level estimates unless there is sufficient contradictory information. Secondly, since any prediction of future (dog) behaviour will entail uncertainty, a Bayesian approach is attractive because it allows the quantification of uncertainty at all levels of analysis [32, 33]. Understanding the uncertainty around individual-level reaction norms is important for making logical predictions about future behaviour.

2 Material & Methods

2.1 Subjects

Behavioural data on N = 3,263 dogs from Battersea Dogs and Cats Home's longitudinal, observational assessment model were used for analysis. The data concerned all behavioural records of dogs at the shelter during 2014 (including those arriving in 2013 or departing in 2015), filtered to include all dogs: 1) at least 4 months of age (to ensure all dogs were treated similarly under shelter protocols, e.g. vaccinated so eligible for walks outside and kennelled in similar areas), 2) with at least one observation during the first 31 days since arrival at the shelter, and 3) with complete data for demographic variables to be included in the formal analysis (Table 1). Since dogs spent approximately one month at the shelter on average (Table 1), we focused on this period in our analyses (arrival day 0 to day 30). We did not include breed characterisation due to the unreliability of using appearance to attribute breed type to shelter dogs of uncertain heritage [34].

2.2 Shelter environment

Details of the shelter environment have previously been presented in [35]. Briefly, the shelter was composed of three different rehoming centres (Table 1): one large inner-city centre based in London (approximate capacity: 150-200 dogs), a medium-sized suburban/rural centre based in Old Windsor (approximate capacity: 100-150 dogs), and a smaller rural centre in Brands Hatch (approximate capacity: 50 dogs). Dogs considered suitable for adoption were

Table 1: *Demographic variables of dogs in the sample analysed. Mean and standard deviation (SD) or the number of dogs by category (N) are displayed.*

Demographic variable	**Mean (SD) / N**
Number of observations per dog	5.9 (3.7)
Days spent at the shelter	25.8 (35.0)
Age (years; all at least 4 months old)	3.7 (3.0)
Weight (kg)	18.9 (10.2)
Source: gift / stray / return	1950 / 1122 / 191
Rehoming centre: London / Old Windsor / Brands Hatch	1873 / 951 / 439
Females / males	1396 / 1867
Neutered: before arrival / at shelter / not / undetermined	1043 / 1281 / 747 / 192

housed in indoor kennels (typically about 4m x 2m, with a shelf and bedding alcove; see also [36]. Most dogs were housed individually, and given daily access to an indoor run behind their kennel. Feeding, exercising and kennel cleaning were performed by a relatively stable group of staff members. Dogs received water ad libitum and two meals daily according to veterinary recommendations. Sensory variety was introduced daily (e.g. toys, essential oils, classical music, access to quiet 'chill-out' rooms). Regular work hours were from 0800 h to 1700 h each day, with public visitation from 1000 h to 1600 h. Unless deemed unsafe, dogs were socialised with staff and/or volunteers daily.

2.3 Data collection

The observational assessment implemented at the shelter included observations of dogs by trained shelter employees in different, everyday contexts, each with its own ethogram of possible behaviours. Shortly after dogs were observed in relevant contexts, employees entered observations into a custom, online platform using computers located in different housing areas. Each behaviour within a context had its own code. Previously, we have reported on aggressive behaviour across contexts [35]. Here, we focus on variation in behaviour in one of the most important contexts, 'Interactions with unfamiliar people', which pertained to how dogs reacted when people with whom they had never interacted before approached, made eye contact, spoke to and/or attempted to make physical contact with them. For the most part, this context occurred outside of the kennel, but it could also occur if an unfamiliar person entered the kennel. Observations could be recorded by an employee meeting an unfamiliar dog, or by an employee observing a dog meeting an unfamiliar person.

Behavioural observations in the 'Interactions with unfamiliar people' context were recorded using a 13-code ethogram (Table 2). Each behavioural code was subjectively labelled and generally defined, providing a balance between behavioural rating and behavioural coding methodologies. The ethogram represented a scale of behavioural problem severity and assumed adoptability (higher codes indicating higher severity of problematic behaviour/lower sociability), reflected by grouping the 13 codes further into green, amber and red codes (Table 2). Green behaviours posed no problems for adoption, amber behaviours suggested dogs may require some training to facilitate successful adoption but did not pose a danger to people or other dogs, and red behaviours suggested dogs needed training or behavioural modification to facilitate successful adoption and could pose a risk to people or other dogs. A dog's suitability for adoption was, however, based on multiple behavioural observations over a number of days. When registering an observation, the employee selected the highest code in the ethogram that was observed on that occasion (i.e. the most severe level of problematic behaviour was given priority). There were periods when a dog could receive no entries for the context for several days but other times when multiple observations were recorded on the same day, usually when a previous observation was followed by a more serious behavioural event. In these instances, and in keeping with the shelter protocol, we retained the highest (i.e. most severe) behavioural code registered for the context that day. When the behaviours were the same, only one record was retained for that day. This resulted in an average of 5.9 (SD = 3.7) records per dog on responses during interactions with unfamiliar people while at the shelter. For dogs with more than one record, the average number of days between records was 2.8 (SD = 2.2).

Table 2: *Ethogram of behavioural codes used to record observations of interactions with unfamiliar people, and their percent prevalence in the sample. Behaviour labels followed by + indicate a more intense form of the behaviour with the same name without a +.*

Behaviour	Colour	%	Definition
1: Friendly	Green	63.5	Dog initiates interactions with people in an appropriate social manner.
2: Excitable	Green	14.2	Animated interaction with an enthusiastic attitude, showing behaviours such as jumping up, mouthing, an inability to stand still, and/or playful behaviour towards people.
3: Independent	Green	4.1	Does not actively seek interaction, although relaxed in the presence of people
4: Submissive	Green	4.6	Appeasing and/or nervous behaviours, including a low body posture, rolling over and other calming signals.
5: Uncomfortable avoids	Amber	5.4	Tense and stiff posture, and/or shows anxious behaviours (e.g. displacement behaviours) while trying to move away from the person.
6: Submissive +	Amber	0.2	High intensity of submissive behaviours such as submissive urination, a reluctance to move, or is frequently overwhelmed by the interaction.
7: Uncomfortable static	Amber	0.8	Tense and stiff posture, and/or shows anxious behaviour (potentially showing displacement behaviours) but doesn?t move away from the person.
8: Stressed	Amber	0.5	High frequency/intensity of stress behaviours, which may include dribbling, stereotypic behaviours, stress vocalisations, constant shedding, trembling, and destructive behaviours.
9: Reacts to people non-aggressive	Amber	2.4	Barks, whines, howls and/or play growls when seeing/meeting people, potentially pulling or lunging towards them.
10: Uncomfortable approaches	Amber	0.7	Tense and stiff posture, and/or shows anxious behaviour (potentially showing displacement behaviours) and approaches the person.
11: Overstimulated	Red	0.8	High intensity of excitable behaviour, including grabbing, body barging, and nipping.
12: Uncomfortable static +	Red	0.1	Body freezes (the body goes suddenly and completely still) in response to an interaction with a person.
13: Reacts to people aggressive	Red	2.8	Growls, snarls, shows teeth and/or snaps when seeing/meeting people, potentially pulling or lunging towards them.

2.4 Validity & inter-rater reliability

Inter-rater reliability and the validity of the assessment methodology were evaluated using data from a larger research project at the shelter. Videos depicting different behaviours in different contexts were filmed by canine behaviourists working at the shelter, who subsequently organised video coding sessions with 93 staff members (each session with about 5 - 10 participants) across rehoming centres [35]. The authors were blind to the videos and administration of video coding sessions. The staff members were shown 14 videos (each about 30 s long) depicting randomly-selected behaviours, two from each of seven different assessment contexts (presented in a pseudo-random order, the same for all participants). Directly after watching each video, they individually recorded (on a paper response form) which ethogram code best described the behaviour observed in each context. Two videos depicted behaviour during interactions with people (familiar versus unfamiliar not differentiated), one demonstrating *Reacts to people aggressive* and the other *Reacts to people non-aggressive* (Table 2). Below, we present the inter-rater reliabilities and the percentage of people who chose the correct behaviour and colour category for these two videos in particular, but also the averaged results across the 14 videos, since there was some redundancy between ethogram scales across contexts.

2.5 Statistical analyses

All data analysis was conducted in R version 3.3.2 [37].

2.5.1 Validity & inter-rater reliability

Validity was assessed by calculating the percentage of people answering with the correct ethogram code/code colour for each video. Inter-rater reliability was calculated for each video using the consensus statistic [38] in the R package *agrmt* [39], which is based on Shannon entropy and assesses the amount of agreement in ordered categorical responses. A value of 0 implies complete disagreement (i.e. responses equally split between the lowest and highest ordinal categories, respectively) and a value of 1 indicates complete agreement (i.e. all responses in a single category). For the consensus statistic, 95% confidence intervals (CIs) were obtained using 10,000 non-parametric bootstrap samples. The confidence intervals were subsequently compared to 95% CIs of 10,000 bootstrap sample statistics from a null distribution, which was created by: 1) selecting the range of unique answers given for a particular video and 2) taking 10,000 samples of the same size as the real data, where each answer had equal probability of being chosen. Thus, the null distribution represented a population with a realistic range of answers, but had no clear consensus about which category best described the behaviour. When the null and real consensus statistics' 95% CIs did not overlap, we inferred statistically significant consensus among participants.

2.5.2 Hierarchical Bayesian ordinal probit model

The distribution of ethogram categories was heavily skewed in favour of the green codes (Table 2), particularly the first *Friendly* category. Since some categories were chosen particularly infrequently, we aggregated the raw responses into a 6-category scale: 1) *Friendly*,

2) *Excitable*, 3) *Independent*, 4) *Submissive*, 5) *Amber codes*, 6) *Red codes*. This aggregated scale retained the main variation in the data and simplified the data interpretation. We analysed the data using a Bayesian ordinal probit model (described in [32, 40]), but extended to integrate the hierarchical structure of the data, including heteroscedastic residual standard deviations to quantify predictability for each dog (for related models, see [31, 41, 42]). The ordinal probit model, also known as the cumulative or thresholded normal model, is motivated by a latent variable interpretation of the ordinal scale. That is, an ordinal dependent variable, Y, with categories K_j, from $j = 1$ to J, is a realisation of an underlying continuous variable divided into thresholds, θ_c, for $c = 1$ to $J - 1$. Under the probit model, the probability of each ordinal category is equal to its area under the cumulative normal distribution, ϕ, with mean, μ, SD σ and thresholds θ_c:

$$Prob(Y = K|\mu, \sigma, \theta_c) = \phi[\frac{\theta_c - \mu}{\sigma}] - \phi[\frac{\theta_{c-1} - \mu}{\sigma}] \tag{1}$$

For the first and last categories, this simplifies to $\phi[(\theta_c - \mu)/\sigma]$ and $1 - \phi[(\theta_{c-1} - \mu)/\sigma]$, respectively. As such, the latent scale extends from $\pm\infty$. Here, the ordinal dependent variable was a realisation of the hypothesised continuum of 'sociability when meeting unfamiliar people', with 6 categories and 5 threshold parameters. While ordinal regression models usually fix the mean and SD of the latent scale to 0 and 1 and estimate the threshold parameters, we fixed the first and last thresholds to 1.5 and 5.5 respectively, allowing for the remaining thresholds, and the mean and SD, to be estimated from the data. As explained by Kruschke [32], this allows for the results to be interpretable with respect to the ordinal scale. We present the results using both the predicted probabilities of ordinal sociability codes and estimates on the latent, unobserved scale assumed to generate the ordinal responses.

2.5.3 Hierarchical structure

To model inter- and intra-individual variation, a hierarchical structure for both the mean and SD was specified. That is, parameters were included for both group-level and dog-level effects. The mean model, describing the predicted pattern of behaviour across days on the latent scale, y^*, for observation i from dog j, was modelled as:

$$y^*_{ij} = \beta_0 + \nu_{0j} + \sum_{p=1}^{P} \beta_{p0} x_{pj} + (\beta_1 + \nu_{1j} + \sum_{p=1}^{P} \beta_{p1} x_{pj}) day_{ij} + (\beta_2 + \nu_{2j} + \sum_{p=1}^{P} \beta_{p2} x_{pj}) day^2_{ij} \tag{2}$$

Equation 2 expresses the longitudinal pattern of behaviour as a function of i) a group-level intercept the same for all dogs, β_0, and the deviation from the group-level intercept for each dog, ν_{0j}, ii) a linear effect of day since arrival, β_1, and each dog's deviation, ν_{1j}, and iii) a quadratic effect of day since arrival, β_2, and each dog's deviation, ν_{2j}. A quadratic effect was chosen based on preliminary plots of the data at group-level and at the individual-level, although we also compared the model's predictive accuracy with simpler models (described below). Day since arrival was standardised, meaning that the intercepts reflected the behaviour on the average day since arrival across dogs (approximately day 8). The three dog-level parameters, ν_j, correspond to personality and linear and quadratic plasticity parameters, respectively. The terms $\sum_{p=1}^{P} \beta_p x_{pj}$ denote the effect of P dog-level predictor variables (x_p), included to explain variance between dog-level intercepts and slopes. These included: the number of observations for each dog, the number of days dogs spent at the shelter controlling for the number of observations (i.e. the residuals from a linear regression of total number of days spent at the shelter on the number of observations), average age while at the shelter, average weight at the shelter, sex, neuter status, source type, and rehoming centre (Table 1). For neuter status, we did not make comparisons between the 'undetermined' category and other categories. The primary goal of including these predictor variables was to obtain estimates of individual differences conditional on relevant inter-individual differences variables, since the data were observational.

The SD model was:

$$\sigma = \exp(\delta + \nu_{3j} + \sum_{p=1}^{P} \beta_{p3} x_{pj}) \tag{3}$$

Equation 3 models the SD of the latent scale by its own regression, with group-level SD intercept, δ, the deviation for each dog from the group-level SD intercept, ν_{3j}, and predictor variables, $\sum_{p=1}^{P} \beta_{p3} x_{pj}$, as in the mean model (equation 2). The SDs across dogs were assumed to approximately follow a log-normal distribution, with $ln(\sigma)$ approximately normally distributed (hence the exponential inverse-link function). The parameter ν_{3j} corresponds to each dog's residual SD or predictability.

All four dog-level parameters were assumed to be multivariate normally distributed with means 0 and variance-covariance matrix $\boldsymbol{\Sigma_\nu}$ estimated from the data:

$$\boldsymbol{\Sigma_\nu} = \begin{bmatrix} \tau_{\nu_0}^2 & \rho_{\nu_0}\tau_{\nu_0}\tau_{\nu_1} & \rho_{\nu_0}\tau_{\nu_0}\tau_{\nu_2} & \rho_{\nu_0}\tau_{\nu_0}\tau_{\nu_3} \\ \cdots & \tau_{\nu_1}^2 & \rho_{\nu_1}\tau_{\nu_1}\tau_{\nu_2} & \rho_{\nu_1}\tau_{\nu_1}\tau_{\nu_3} \\ \cdots & \cdots & \tau_{\nu_2}^2 & \rho_{\nu_2}\tau_{\nu_2}\tau_{\nu_3} \\ \cdots & \cdots & \cdots & \tau_{\nu_3}^2 \end{bmatrix} \tag{4}$$

The diagonal elements are the variances of the dog-level intercepts, linear slopes, quadratic slopes and residual SDs, respectively, while the covariances fill the off-diagonal elements (only the upper triangle shown), where ρ is the correlation coefficient. In the results, we report $\tau_{\nu 3}$ (the SD of dog-level residual SDs) on the original scale, rather than the log-transformed scale, using $\sqrt{e^{2\delta+\tau_{\nu 3}^2} e^{\tau_{\nu 3}^2} - 1}$. Likewise, δ was transformed to the median of the original scale by e^δ.

To summarise the amount of behavioural variation explained by differences between individuals, referred to as repeatability in the personality literature [1], we calculated the intra-class correlation coefficient (ICC). Since the model includes both intercepts and slopes varying by dog, the ICC is a function of both linear and quadratic effects of day since arrival. The ICC for day i, assuming individuals with the same residual variance (i.e. using the median of the log-normal residual SD), was calculated as:

$$ICC_i = \frac{\tau_{\nu_0}^2 + 2Cov_{\nu_0,\nu_1} Day_i^2 + 2Cov_{\nu_0,\nu_2} Day_i^2 + \tau_{\nu_2}^2 Day_i^4 + 2Cov_{\nu_1,\nu_2} Day_i^3}{numerator + e^\delta} \tag{5}$$

Equation 5 is an extension of the intra-class correlation calculated from mixed-effect models with a random intercept only [43] to include the variance parameters for, and covariances between, the linear and quadratic effects of day, which were evaluated at specific days of interest. We calculated the ICC for values of -1, 0 and 1 on the standardised day scale, corresponding to approximately the arrival day (day 0), day 8, and day 15. This provided a representative spread of days for most of the dogs in the sample, since there were fewer data available for later days which could lead to inflation of inter-individual differences. To inspect how much the rank-order differences between dogs changed from arrival day compared to later days, we calculated the 'cross-environmental' correlations [44] between the same days as the ICC. Although correlations between intercept and slope parameters provide some indication of the amount of crossing between individuals' reaction norms through time, the cross-environmental correlation offers a more direct measure of rank-order change across particular environments, where 'days since arrival' is, here, a special case of differing 'environments' [44]. The cross-environmental covariance matrix, $\mathbf{\Omega}$, between the three focal days was calculated as:

$$\mathbf{\Omega} = \mathbf{\Psi}\boldsymbol{K}\mathbf{\Psi}^{\mathsf{T}} \tag{6}$$

In equation 6, $\boldsymbol{K}$ represents the variance-covariance matrix of the dog-level intercepts and (linear and quadratic) slopes, and $\mathbf{\Psi}$ is a three-by-three matrix with a column vector of 1s and two column vectors containing -1, 0 and 1 (defining the days for the cross-environmental correlations). Once defined, $\mathbf{\Omega}$ was scaled to a correlation matrix. Finally, to summarise the degree of individual differences in predictability, we calculated the 'coefficient of variation for predictability' as $\sqrt{e^{\tau^2_{\nu_3}} - 1}$ following Cleasby *et al.* [5].

2.5.4 Prior distributions

We chose prior distributions that were either weakly informative (i.e. specified a realistic range of parameter values) for computational efficiency, or weakly regularising to prioritise conservative inference. The prior for the overall intercept, β_0, was $Normal(\bar{y}, 5)$, where $\bar{y}$ is the arithmetic mean of the ordinal data. The linear and quadratic slope parameters, β_1 and β_2, were given $Normal(0, 1)$ priors. Coefficients for the dog-level predictor variables, β_k, were given $Normal(0, \sigma_{\beta_p})$ priors, where σ_{β_p} was a shared SD across predictor variables, which had in turn a half-Cauchy hyperprior with mode 0 and shape parameter 2, *half-Cauchy*$(0, 2)$. Using a shared SD imposes shrinkage on the regression coefficients for conservative inference: when most regression coefficients are near zero, then estimates for other regression coefficients are also pulled towards zero (e.g. [32]). The prior for the overall log-transformed residual SD, δ, was $Normal(0, 1)$. The covariance matrix of the random effects was parameterised as a Cholesky decomposition of the correlation matrix (see [45] for more details), where the SDs had *half-Cauchy*$(0, 2)$ priors and the correlation matrix had a LKJ prior distribution [46] with shape parameter η set to 2.

2.5.5 Model selection & computation

We compared the full model explained above to five simpler models. Starting with the full model, the alternative models included: i) parameters quantifying personality and quadratic and linear plasticity only; ii) parameters quantifying personality and linear plasticity only, with a fixed quadratic effect of day since arrival; iii) parameters quantifying personality only, with fixed linear and quadratic effects of day since arrival; iv) parameters quantifying personality only, with a fixed linear effect of day since arrival; and v) a generalised linear regression with no dog-varying parameters and a linear fixed effect for day since arrival (Figure 2). Models were compared by calculating the widely applicable information criterion (WAIC; [47]) following McElreath [33] (see the R script file). The WAIC is a fully Bayesian information criterion that indicates a model's *out-of-sample* predictive accuracy relative to other plausible models while accounting for model complexity. Thus, WAIC guards against both under- and over-fitting to the data (unlike measures of purely in-sample fit, e.g. R^2).

Models were computed using the probabilistic programming language Stan [45] using the *RStan* package [48] version 2.15.1, which employs Markov chain Monte Carlo estimation using Hamiltonian Monte Carlo (see the R script file and Stan code for full details). We ran four chains of 5,000 iterations each, discarding the first 2,500 iterations of each chain as warm-up, and setting thinning to 1. Convergence was assessed visually using trace plots to ensure chains were well mixed, numerically using the Gelman-Rubin statistic (values close to 1 and < 1.05 indicating convergence) and by inspecting the effective sample size of each parameter. We also used graphical posterior predictive checks to assess model predictions against the raw data, including 'counterfactual' predictions [33] to inspect how dogs would be predicted to behave across the first month of being in the shelter regardless of their actual number of observations or length of stay at the shelter. To summarise parameter values, we calculated mean (denoted β) and 95% highest density intervals (HDIs), the 95% most probable values for each parameter (using functions in the *rethinking* package; [33]). For comparing levels of categorical variables, the 95% HDI of their differences were calculated (i.e. the differences between the coefficients at each step in the MCMC chain, denoted β_{diff}). When the 95% HDI of predictor variables surpassed zero, a credible effect was inferred.

3 Results

3.1 Inter-rater reliability & validity

For the two videos depicting interactions with people, consensus was 0.75 (95% CI: 0.66, 0.84) for the video showing an example of *Reacts to people non-aggressive* and 0.77 (95% CI: 0.74, 0.81) for the example of *Reacts to people aggressive*, respectively. Neither did these results overlap with the null distributions (see Supplementary Material Table S1), indicating significant inter-rater reliability. For the video showing *Reacts to people non-aggressive*, 77% chose the correct code and 83% a code of the correct colour category (amber), and, as previously reported by [35], 52% chose the correct code for the video showing *Reacts to people aggressive* and 55% chose a code of the correct colour category (red; 42% chose the amber code *Reacts to people non-aggressive* instead). Across all assessment context videos, the average consensus was 0.71 and participants chose the correct ethogram category 66% of the time while 78% of answers were a category of the correct ethogram colour.

3.2 Hierarchical ordinal probit model

The full model had the best out-of-sample predictive accuracy, with the inclusion of heterogeneous residual SDs among dogs improving model fit by over 1,500 WAIC points compared to the second most plausible model (Alternative 1 in Figure 2). In general, models that included more parameters to describe personality, plasticity and predictability, and models with a quadratic effect of day, had better out-of-sample predictive accuracy, despite the added complexity brought by additional parameters.

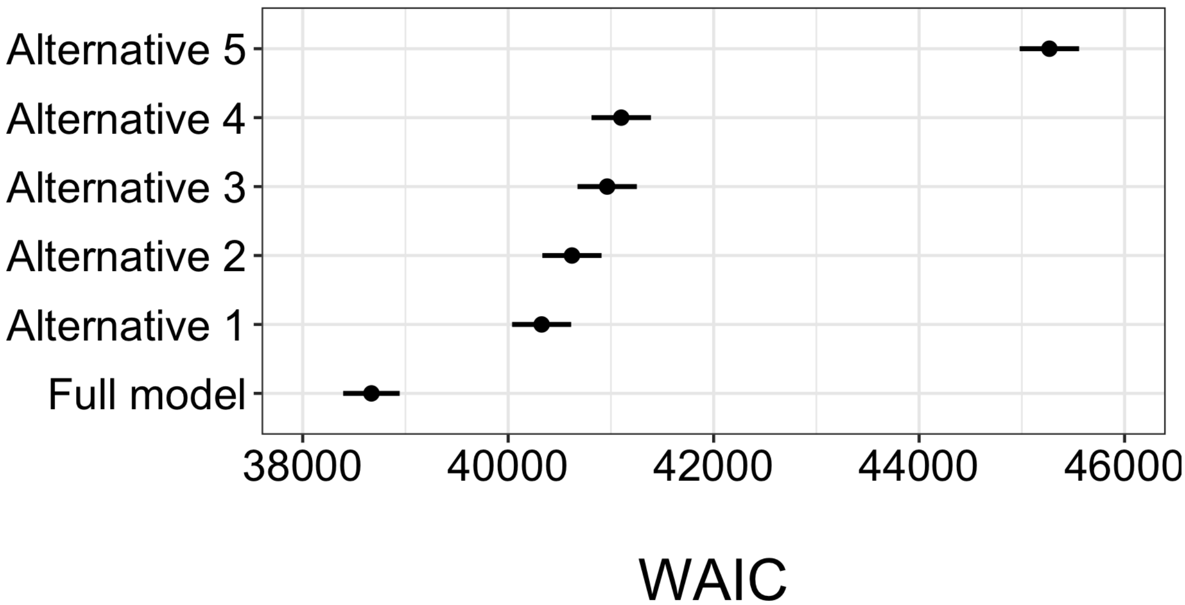
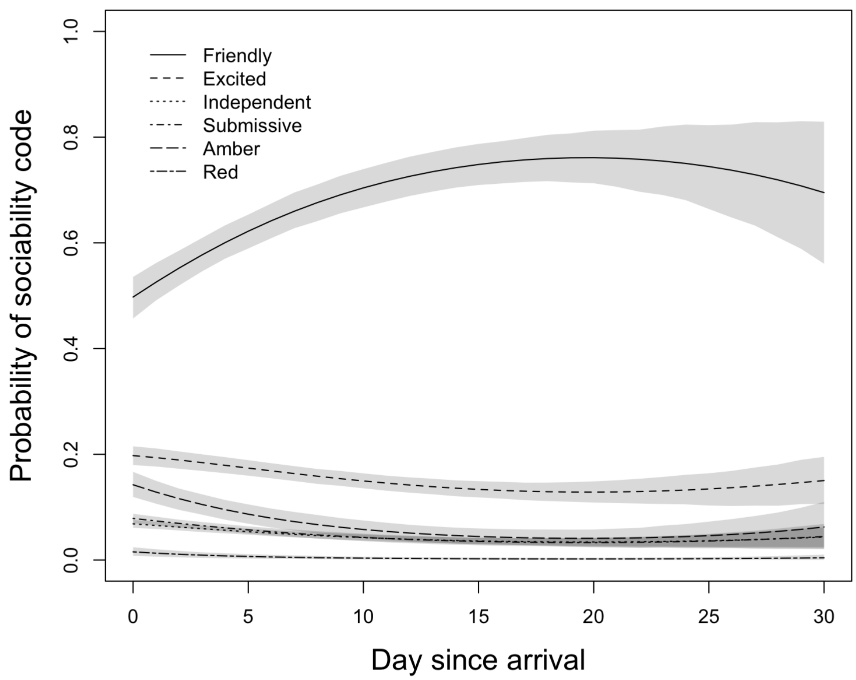


Figure 2: *Out-of-sample predictive accuracy (lower is better) for each model (described in text section 2.5.5) measured by the widely applicable information criterion (WAIC). Black points denote the WAIC estimate and horizontal lines show WAIC estimates ± standard error. Mean ± standard error: full model = 38669 ± 275; alternative 1 = 40326 ± 288; alternative 2 = 40621 ± 288; alternative 3 = 40963 ± 289; alternative 4 = 41100 ± 289; alternative 5 = 45268 ± 289.*

At the group-level, the *Friendly* code (Table 2) was most probable overall and was estimated to increase in probability across days since arrival, while the remaining sociability codes either decreased or stayed at low probabilities (Figure 3a), reflecting the raw data. On the latent sociability scale (Figure 3b), the group-level intercept parameter on the average day was 0.68 (95% HDI: 0.51, 0.86). A one SD increase in the number of days since arrival was associated with a -0.63 unit (95% HDI: -0.77, -0.50) change on the latent scale on average (i.e. reflecting increasing sociability), and the group-level quadratic slope was positive ($\beta = 0.20$, 95% HDI: 0.10, 0.30), reflecting a quicker rate of change in sociability earlier after arrival to the shelter than later (i.e. a concave down parabola). There was a slight increase in the quadratic curve towards the end of the one-month period, although there were fewer behavioural observations at this point and so greater uncertainty about the exact shape of the curve, resulting in estimates being pulled closer to those of the intercepts. The group-level residual standard deviation had a median of 1.84 (95% HDI: 1.67, 2.02).

At the individual level, heterogeneity existed in behavioural trajectories across days since arrival (Figure 3b). The SDs of dog-varying parameters were: i) intercepts: 1.29 (95% HDI: 1.18, 1.41; Figure 4a), ii) linear slopes: 0.56 (95% HDI: 0.47, 0.65; Figure 4b), iii) quadratic

(a)



(b)

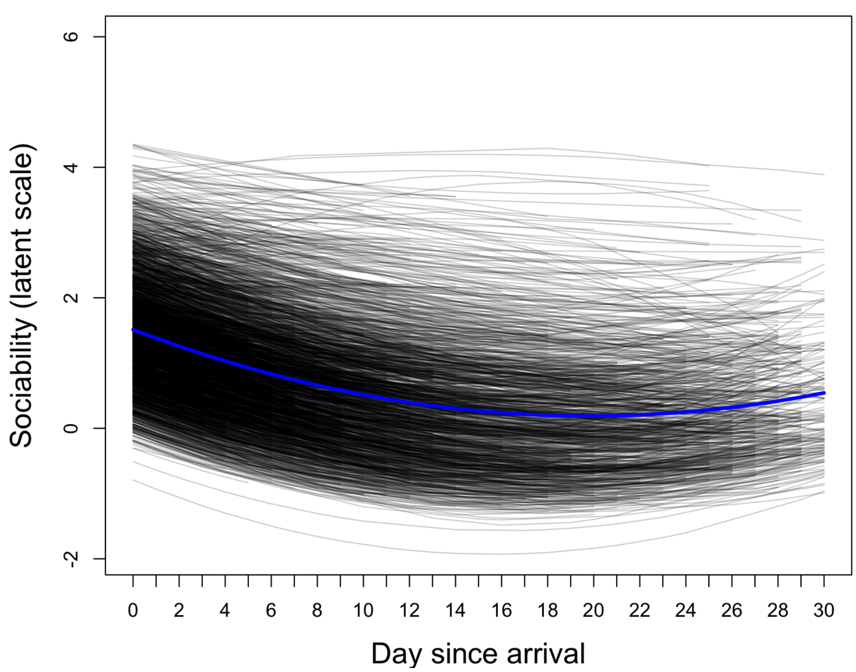


Figure 3: (a) *Predicted probabilities (posterior means = black lines; 95% highest density intervals = shaded areas) of different sociability codes across days since arrival.* (b) *Posterior mean behavioural trajectories on the latent scale (ranging from $\pm\infty$) at the group-level (blue line) and for each individual (black lines), where higher values indicate lower sociability.*

slopes: 0.28 (95% HDI: 0.20, 0.35; Figure 4c), and iv) residual SDs: 1.39 (95% HDI: 1.22, 1.58; Figure 4d). There was also large uncertainty in individual-level estimates. Figure 5 displays counterfactual model predictions for twenty randomly-sampled dogs. Uncertainty in reaction norm estimates, illustrated by the width of the 95% HDIs (dashed black lines), was greatest when data were sparse (e.g. towards the end of the one-month study period). Hierarchical shrinkage meant that individuals with observations of less sociable responses, or individuals with few behavioural observations, tended to have model predictions pulled towards the overall mean. Note that regression lines depict values on the latent scale predicted to generate observations on the ordinal scale, and so may not clearly fit the ordinal data points. The coefficient of variation for predictability was 0.64 (95% HDI: 0.58, 0.70). Individuals with the five highest and lowest residual SD estimates are shown in Figure 6.

Dog-varying intercepts positively correlated with linear slope parameters (ρ = 0.38, 95% HDI: 0.24, 0.50) and negatively correlated with quadratic slope parameters (ρ = -0.54, 95% HDI: -0.68, -0.39), and linear and quadratic slopes had a negative correlation (ρ = -0.75, 95% HDI: -0.88, -0.59), indicating that less sociable individuals (with higher scores on the ordinal scale) had flatter reaction norms on average. Dog-varying residual SDs had a correlation with the intercept parameters of approximately zero (ρ = 0.00, 95% HDI: -0.10, 0.10) but were negatively correlated with the linear slope parameters (ρ = -0.37, 95% HDI: -0.51, -0.22) and positively correlated with the quadratic slopes (ρ = 0.24, 95% HDI: 0.05, 0.42), indicating that dogs with greater residual SDs were predicted to change the most across days since arrival.

The ICC by day increased through time, ranging from 0.18 (95% HDI: 0.11, 0.24) on day 0 (arrival day) to 0.33 (95% HDI: 0.28, 0.38) on day 8 to 0.35 (95% HDI: 0.30, 0.41) on day 15. The cross-environmental correlation between days 0 and 8 was 0.79 (95% HDI: 0.70, 0.88), between days 0 and 15 was 0.51 (95% HDI: 0.35, 0.68), and between days 8 and 15 was 0.95 (95% HDI: 0.93, 0.97).

A one SD increase in the number of observations was associated with higher intercepts (β = 0.12; 95% HDI: 0.03, 0.21; see Supplementary Material Table S2) and higher residual SDs (β = 0.06, 95% HDI: 0.02, 0.10). Increasing age by one SD was associated with lower intercepts (β= -0.61, 95% HDI: -0.70, -0.51), steeper linear slopes (β = -0.20, 95% HDI: -0.27, -0.13), a stronger quadratic curve (β= 0.07, 95% HDI: 0.03, 0.12), and larger residual SDs (β = 0.05, 95% HDI: 0.01, 0.09). Increasing weight by one SD was associated with shallower quadratic curves (β = -0.05, 95% HDI: -0.09, -0.01). No credible effect of sex was observed on personality, plasticity nor predictability. Gift dogs had larger intercepts than returned dogs (β_{diff} = 0.28, 95% HDI: 0.04, 0.52) and stray dogs (β_{diff} = 0.33, 95% HDI: 0.15, 0.50), as well as steeper linear slopes (β_{diff} = -0.25, 95% HDI: -0.38, -0.13) and higher residual SDs than stray dogs (β_{diff} = 0.10, 95% HDI: 0.02, 0.18). Dogs at the large rehoming centre had steeper linear slopes (β_{diff} = -0.70, 95% HDI: -0.84, -0.56) and stronger quadratic curves (β_{diff} = 0.35, 95% HDI: 0.26, 0.45) than dogs at the medium rehoming centre, and lower intercept parameters (β_{diff} = -0.30, 95% HDI: -0.50, -0.09) and steeper linear slopes (β_{diff} = -0.22, 95% HDI: -0.38, -0.06) than dogs at the small rehoming centre. Compared to dogs at the small rehoming centre, dogs at the medium centre had lower intercepts (β_{diff}= -0.25, 95% HDI: -0.48, -0.01), and shallower linear (β_{diff} = 0.48, 95% HDI: 0.30, 0.66) and quadratic slopes (β_{diff} = -0.34, 95% HDI: -0.46, -0.22). Dogs already neutered before arrival to the shelter had lower intercepts (β_{diff} = -0.54, 95% HDI: -1.07, -0.03) and lower residual

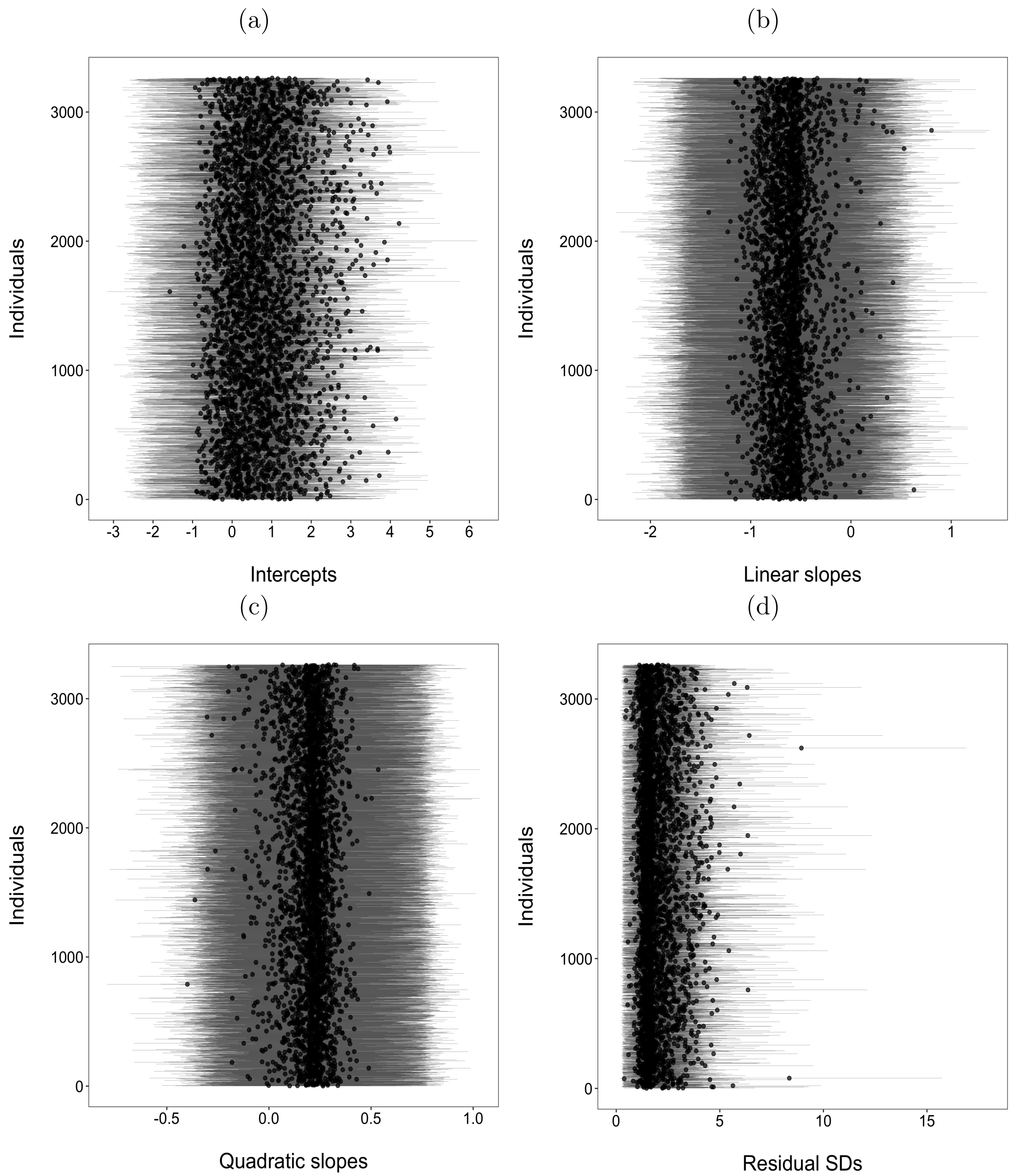


Figure 4: *Posterior means (black dots) and 95% highest density intervals (grey vertical lines) for each dogs' (a) intercept, (b) linear slope, (c) quadratic slope, and (d) residual SD parameter.*

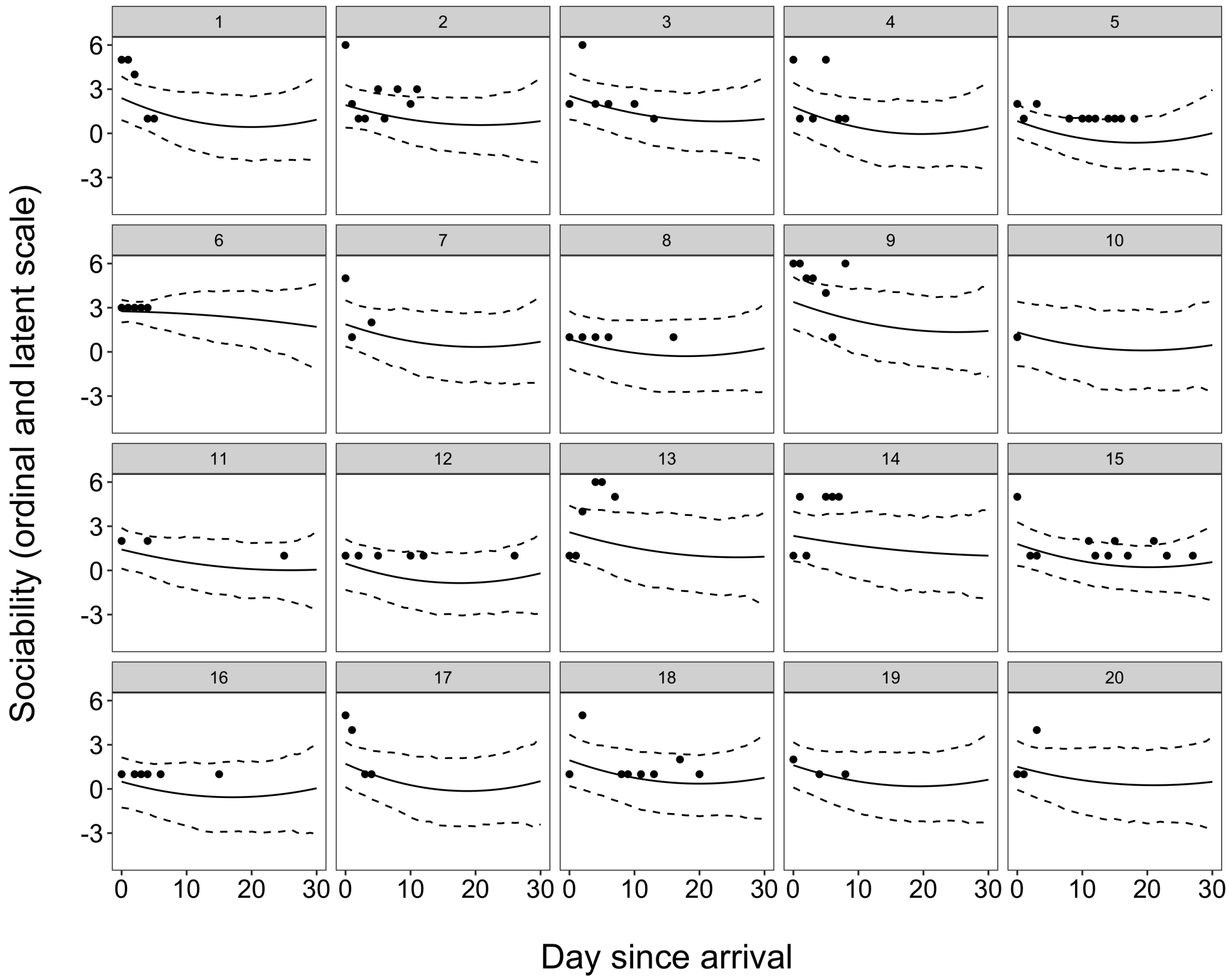


Figure 5: *Predicted reaction norms ('counterfactual' plots) for twenty randomly-selected dogs. Black points show raw data on the ordinal scale, where higher values indicate lower sociability, and solid and dashed lines illustrate posterior means and 95% highest density intervals (HDI). When data were sparse, there was increased uncertainty in model predictions. Due to hierarchical shrinkage, individual dogs' model predictions were pulled towards the group-level mean, particularly for those dogs showing higher behavioural codes (where higher values indicate lower sociability).*

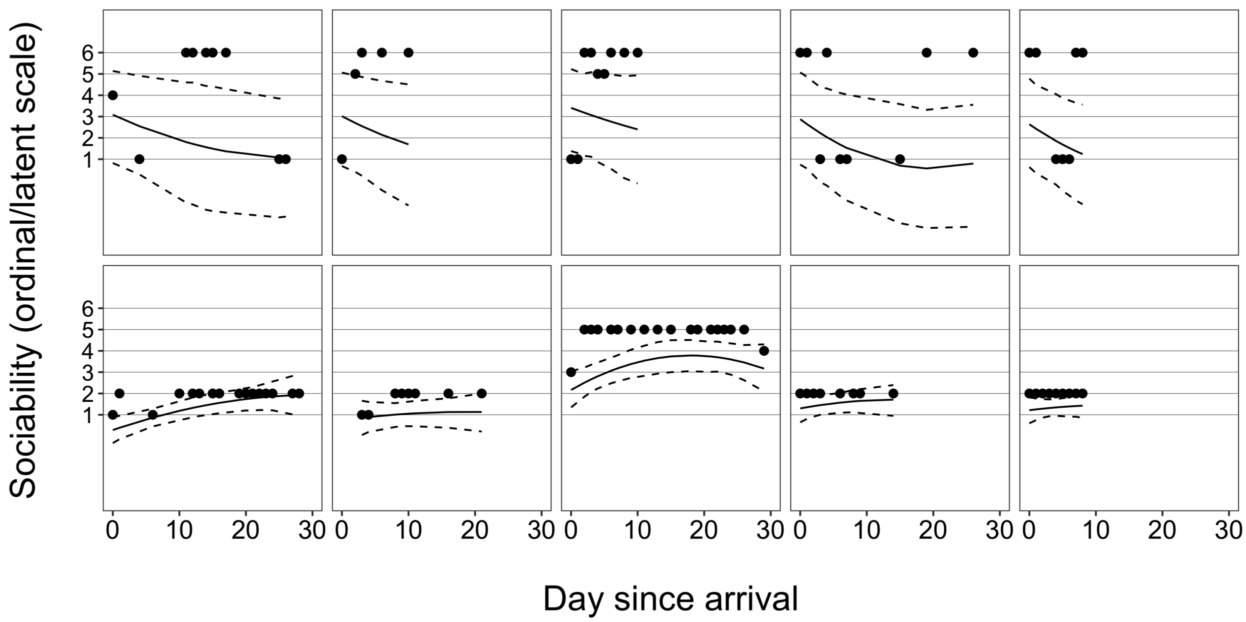


Figure 6: *Reaction norms (posterior means = solid black lines; 95% highest density intervals = dashed black lines) for individuals with the five highest (top row) and five lowest (bottom row) residual SDs. Black points represent raw data on the ordinal scale.*

SDs (β_{diff} = -0.53, 95% HDI: -0.85, -0.22) than dogs not neutered, but higher intercepts (β_{diff} = 0.20, 95% HDI: 0.03, 0.37) and higher residual SDs (β_{diff} = 0.10, 95% HDI: 0.02, 0.19) than those neutered whilst at the shelter. Unneutered dogs had higher intercepts (β_{diff} = 0.74, 95% HDI: 0.20, 1.26) and higher residual SDs (β_{diff} = 0.63, 95% HDI: 0.30, 0.92) than dogs neutered at the shelter.

4 Discussion

This study applied the framework of behavioural reaction norms to quantify inter- and intra-individual differences in shelter dog behaviour during interactions with unfamiliar people. This is the first study to systematically analyse behavioural data from a longitudinal, observational assessment of shelter dogs. Dogs demonstrated substantial individual differences in personality, plasticity and predictability, which were not well described by simply investigating how dogs behaved on average. In particular, accounting for individual differences in predictability, or the short-term, day-to-day fluctuations in behaviour, resulted in significant improvement in the analyses (Figure 2). Modelling dogs' longitudinal behaviour also demonstrated behavioural repeatability increased with days since arrival, and that while individuals maintained rank-order differences in sociability across smaller periods (e.g. one week), rank-order differences were only moderately maintained between arrival to the shelter and day 15. The results highlight the importance of adopting observational and longitudinal assessments of shelter dog behaviour [24], provide a method by which to analyse longitudinal data commensurate with other work in animal behaviour, and identify previously unconsidered behavioural measures that could be used to improve the predictive validity of behavioural assessments in dogs.

4.1 Average behaviour

At the group-level, dogs' reactions to meeting unfamiliar people were predominantly coded as *Friendly* (Figure 3a), described as 'Dog initiates interactions in an appropriate social manner'. Although this definition is broad, it represents a functional qualitative characterisation of behaviour suitable for the purposes of the shelter when coding behavioural interactions, and its generality may partly explain why it was the most prevalent category. The results are consistent with findings that behaviours indicative of poor welfare and/or difficulty of managing (e.g. aggression) are relatively infrequent even in the shelter environment [22, 26]. The change of behaviour across days since arrival was characterised by an increase in the *Friendly* code and a decrease in other behavioural codes (Figure 3a). Furthermore, the positive quadratic effect of day since arrival on sociability illustrates that the rate of behavioural change was not constant across days, being quickest earlier after arrival (Figure 3b). The range of behavioural change at the group-level was, nevertheless, still concentrated around the lowest behavioural codes, *Friendly* and *Excitable*.

Previous studies provide conflicting evidence regarding how shelter dogs adapt to the kennel environment over time, including behavioural and physiological profiles indicative of both positive and negative welfare [26]. Whereas some authors report decreases in the prevalence of some stress- and/or fear related behaviour with time [27, 49], others have reported either no change or an increase in behaviours indicative of poor welfare [17, 30]. Of relevance here, Kis *et al.* [17] found that aggression towards unknown people increased over the first two weeks of being at a shelter. Here, aggression was rare (Table 2), and the probability of 'red codes' (which included aggression) decreased with days at the shelter (Figure 3a). A salient difference between the latter study and the one reported here is that Kis *et al.* [17] collected data using a standardised behavioural test consisting of a stranger engaging in a 'threatening approach' towards dogs. By contrast, we used a large data set of behavioural observations recorded after non-standardised, spontaneous interactions between dogs and unfamiliar people. In recording spontaneous interactions, the shelter aimed to elicit behaviour more representative of a dog's typical behaviour outside of the shelter environment than would be seen in a standardised behavioural assessment. Previously, authors have noted that standardised behavioural assessments may induce stress to individuals and inflate the chances of dogs displaying aggression [29], emphasising the need for observational methods of assessment in shelters [24]. While such observational methods are less standardised, they may have greater ecological validity by giving results more representative of how dogs will behave outside of the shelter. Testing the predictive value of observational assessments on behaviour post-adoption is the focus of future research.

4.2 Individual-level variation

When behavioural data are aggregated across individuals, results may provide a poor representation of how individuals in a sample actually behaved. Here, we found heterogeneity in dog behaviour across days since arrival, even after taking into account a number of dog-level predictor variables that could explain inter-individual differences. Variation in individuals' average behaviour across days (i.e. variation in dogs' intercept estimates) illustrated that personality estimates spanned a range of behavioural codes, although model predictions were mostly focused on the green codes (Figure 3b; Table 2). However, whilst there were many records to inform group-level estimates, there were considerably fewer records available for each individual, which resulted in large uncertainty of individual personality parameters (illustrated by wide 95% HDI bars in Figure 4a). Personality variation has been the primary focus of previous analyses of individual differences in dogs, often based on data collected at one time point and usually on a large number of behavioural variables that require reduction into composite or latent variables (e.g. [50–52]). Our results highlight that ranking individuals on personality dimensions from few observations entails substantial uncertainty.

Certain studies on dog personality have explored how personality trait scores change across time periods, such as ontogeny (e.g. [53]) or time at a shelter (e.g. [17]). Such analyses assume, however, that individuals have similar degrees of change through time. If individuals differ in the magnitude or direction of change (i.e. different degrees of plasticity), group-level patterns of change may not capture important individual heterogeneity. In this study, most dogs were likely to show lower behavioural codes/more sociable responses across days since arrival, although the rate of linear and quadratic change differed among dogs. Indeed, some dogs showed a *decrease* in sociability through time (individuals with positive model estimates in Figure 4b), and while most dogs showed greater behavioural change early after arrival, others showed slower behavioural change early after arrival (individuals with negative model estimates in Figure 4c). As with estimates of personality, there was also large uncertainty of plasticity.

Part of the difficulty of estimating reaction norms for heterogeneous data is choosing a function that best describes behavioural change. We used both linear and quadratic effects of day since arrival based on preliminary plots of the data, supported by lower WAIC values compared to a model with just a linear effect of day since arrival (alternative model 3 versus 4 in Figure 2). Low-order polynomial functions were also relatively easy to vary across individuals while maintaining interpretability of the results. Most studies are, nevertheless, constrained to first-order polynomial reaction norms through time due to collecting data at only a few time points [6, 44], and even higher-order polynomial functions may only produce crude representations of data-generating processes [33]. More complex functions (e.g. regression splines), on the other hand, have the disadvantage of being less easily interpretable. By collecting data more intensely, the opportunities to model behavioural reaction norms with biologically-informed functions of contexts and time should improve. For instance, the rise of ecological momentary assessment studies in psychology has allowed greater possibilities in the modelling of behaviour as a dynamic system (e.g. [54, 55]).

Personality and plasticity were correlated, with dogs with less sociable behaviour across days being less plastic. Previous studies have explored the relationship between how individuals behave on average and their degree of behavioural change. David *et al.* [56] found that male golden hamsters (*Mesocricetus auratus*) showing high levels of aggression in a social intruder paradigm were slower in adapting to a delayed-reward paradigm. In practice, the relationship between personality and plasticity is probably context dependent. Betini and Norris [57] found, for instance, that more aggressive male tree swallows (*Tachycineta bicolor*) during nest defence were more plastic in response to variation in temperature, but that plasticity was only advantageous for nonaggressive males and no relationship was present between personality and plasticity in females. The correlation between personality and plasticity indicates a 'fanning out' shape of the reaction norms through time (Figure 3b). Consequently, behavioural repeatability increased as a function of day. The 'cross-environmental' correlation, moreover, indicated that the most sociable dogs on arrival day were not necessarily the most sociable on later days at the shelter. In particular, the correlation between sociability scores on arrival day and day 15 was only moderate, supporting Brommer [44] that the rank-ordering of trait scores is not always reliable. By contrast, the cross-environmental correlation between days 0 and 8, and 8 and day 15 were much stronger. These results suggest that shelters using standardised behavioural assessments would benefit from administering such tests as late as possible after dogs arrive.

Of particular interest was predictability or the variation in dogs' residual SDs. Predictability has received little attention in research on (shelter) dogs although some have posited that dogs may vary in their behavioural consistency (e.g. [13]). Distinguishing between inter- and intra-individual variation, as done here, is key to testing this hypothesis. Modelling residual SDs for each dog resulted in a model with markedly better out-of-sample predictive accuracy (Figure 2). The coefficient of variation for predictability was 0.64 (95% HDI: 0.58, 0.70), which is high compared to other studies in animal behaviour. For instance, Mitchell *et al.* [6] reported a value of 0.43 (95% HDI: 0.36, 0.53) in spontaneous activity measurements of male guppies (*Poecilia reticulata*). Variation in predictability also supports the hypothesis that dogs have varying levels of behavioural consistency. It is important to note, however, that interactions with unfamiliar people at the shelter were likely more heterogeneous than behavioural measures from standardised tests or laboratory environments, which may contribute to greater individual variation in predictability. Moreover, the behavioural data here may have contained more measurement error than more standardised environments. Although shelter employees demonstrated significant inter-rater reliability in video coding sessions, the average proportion of shelter employees who selected the correct behavioural code to describe behaviours seen in videos was only 66%, while 78% chose a video in the correct colour category (green, amber or red). For observational methods in shelters, it is essential to evaluate the reliability and validity of behavioural records since the observational contexts will be less standardised. Defining acceptable standards of reliability and validity is, however, non-trivial and we could not find measures of reliability or validity in any of the previous studies investigating predictability in animals for comparison.

Dogs with higher residual SDs demonstrated steeper linear slopes and greater quadratic curves, indicating that greater plasticity was associated with lower predictability. The costs of plasticity are believed to include greater phenotypic instability, in particular developmental instability [11, 58]. Since more plastic individuals are more responsive to environmental perturbation, a limitation of plasticity may be greater phenotypic fluctuation on finer time scales. However, lower predictability may also confer a benefit to individuals precisely because they are less predictable to con- and hetero-specifics. For instance, Highcock and Carter [59] reported that predictability in behaviour decreases under predation risk in Namibian rock agamas (*Agama planiceps*). No correlation was found here between personality and predictability, similar to findings of Biro and Adriaenssens [2] in mosquitofish (*Gambusia holbrooki*), although correlations were found in agamas [59] and guppies [6].

4.3 Predictors of individual variation

Finally, we found associations between certain predictor variables and personality, plasticity and predictability (Table S2). Our primary reason for including these predictor variables was to obtain more accurate estimates of personality, plasticity and predictability, and we remain cautious about *a posteriori* interpretations of their effects, especially since the theory underlying why individuals may, for example, demonstrate differences in predictability is in its infancy [8]. The reproducibility of a number of the results would, nevertheless, be interesting to confirm in future research. In particular, understanding factors affecting intra-individual change is important since many personality assessments are used to predict an individual's future behaviour, rather than understand inter-individual differences. Here, increasing age was associated with greater plasticity (linear and quadratic change) and lower predictability, although some of the parameters' 95% HDIs were close to zero, indicative of small effects. In great tits (*Parus major*) conversely, plasticity decreased with age [60], whilst in humans, intra-individual variability in reaction times increased with age [61]. Moreover, non-neutered dogs showed lower predictability than neutered dogs, and dogs entering the shelter as gifts (relinquished by their owners) had lower predictability estimates than stray dogs (dogs brought in by local authorities or members of the public after being found without their owners). Although these results can be used to formulate specific hypotheses about behavioural variation, researchers should beware of making generalisations based on inter-individual differences without first assessing the amount of individual-level heterogeneity.

5 Conclusion

We applied the framework of behavioural reactions norms to data from a longitudinal and observational shelter dog behavioural assessment, quantifying inter- and intra-individual behavioural variation in dogs' interactions with unfamiliar people. Overall, shelter dogs were sociable with unfamiliar people and sociability continued to increase with days since arrival to the shelter. At the same time, dogs showed individual differences in personality, plasticity and predictability. Accounting for all of these components substantially improved the analyses, particularly the inclusion of predictability, which suggests that individual differences in day-to-day behavioural variation is an important, yet largely unstudied, component of dog behaviour. Our results also highlight the uncertainty of making predictions on shelter dog behaviour, particularly when the number of behavioural observations is low. For shelters conducting standardised behavioural assessments, assessments are likely best carried out as late as possible, given that rank-order differences between individuals were only moderately related between arrival and at day 15. In conclusion, this study supports moving towards observational and longitudinal assessments of shelter dog behaviour, has demonstrated a Bayesian method by which to analyse longitudinal data on dog behaviour, and suggests that the predictive validity of behavioural assessments in dogs could be improved by systematically accounting for both inter- and intra-individual variation.

6 Ethics statement

Full permission to use the data in this article was provided by Battersea Dogs and Cats Home.

7 Data accessibility

The data, R code and Stan model code to run the analyses and produce the results and figures in this article are available on Github: https://github.com/ConorGoold/GooldNewberry_modelling_shelter_dog_behaviour

8 Competing interests

We declare no competing interests.

9 Author contributions

CG and RCN conceptualised the study. CG obtained the data, conducted the statistical analyses and drafted the initial manuscript. CG and RCN revised the manuscript and wrote the final version.

10 Acknowledgements

The authors are especially grateful to Battersea Dogs and Cats Home for providing the data on their behavioural assessment.

11 Funding statement

CG and RCN are employed by the Norwegian University of Life Sciences. No additional funding was required for this study.

References

[1] Bell AM, Hankison SJ, Laskowski KL. 2009. The repeatability of behaviour: a meta-analysis. *Anim. Behav.* **77**, 771–783. doi: 10.1016/j.anbehav.2008.12.022.

[2] Biro PA, Adriaenssens B. 2013. Predictability as a personality trait: consistent differences in intraindividual behavioral variation. *Am. Nat.* **182**, 621–629. doi: 10.1086/673213.

[3] Dingemanse NJ, Kazem AJN, Réale D, Wright J. 2010. Behavioural reaction norms: animal personality meets individual plasticity. *Trends Ecol. Evol.* **25**, 81–89. doi: 10.1016/j.tree.2009.07.013.

[4] Bridger D, Bonner SJ, Briffa M. 2015. Individual quality and personality: bolder males are less fecund in the hermit crab (*Pagurus bernhardus*). *Proc. R. Soc. B* **282**, 20142492. doi: 10.1098/rspb.2014.2492.

[5] Cleasby IR, Nakagawa S, Schielzeth H. 2015. Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. *Methods Ecol. Evol.* **6**, 27–37. doi: 10.1111/2041-210X.12281.

[6] Mitchell DJ, Fanson BG, Beckmann C, Biro PA. 2016. Towards powerful experimental and statistical approaches to study intraindividual variability in labile traits. *R. Soc. Open Sci.* **3**, 160352. doi: 10.1098/rsos.160352.

[7] Stamps JA, Briffa M, Biro PA. 2012. Unpredictable animals: individual differences in intraindividual variability (IIV). *Anim. Behav.* **83**, 1325–1334. doi: 10.1016/j.anbehav.2012.02.017.

[8] Westneat DF, Wright J, Dingemanse NJ. 2015. The biology hidden inside residual within-individual phenotypic variation. *Biol. Rev.* **90**, 729–743. doi: 10.1111/brv.12131.

[9] DeWitt TJ. 2016. Expanding the phenotypic plasticity paradigm to broader views of trait space and ecological function. *Curr. Zool.* **62**, 463–473. doi: 10.1093/cz/zow085.

[10] Scheiner SM. 2014. The genetics of phenotypic plasticity. XIII. Interactions with developmental instability. *Ecol. Evol.* **4**, 1347–1360. doi: 10.1002/ece3.1039.

[11] Tonsor SJ, Elnaccash TW, Scheiner SM. 2013. Developmental instability is genetically correlated with phenotypic plasticity, constraining heritability, and fitness. *Evolution* **67**, 2923–2935. doi: 10.1111/evo.12175.

[12] Oates AC. 2011. What's all the noise about developmental stochasticity? *Development* **138**, 601–607. doi: 10.1242/dev.059923.

[13] Fratkin JL, Sinn DL, Patall EA, Gosling SD. 2013. Personality consistency in dogs: a meta-analysis. *PLoS ONE* **8**, e54907. doi: 10.1371/journal.pone.0054907.

[14] Wilsson E, Sundgren PE. 1998. Behaviour test for eight-week old puppies—heritabilities of tested behaviour traits and its correspondence to later behaviour. *Appl. Anim. Behav. Sci.* **58**, 151–162. doi: 10.1016/S0168-1591(97)00093-2.

[15] Sinn DL, Gosling SD, Hilliard S. 2010. Personality and performance in military working dogs: reliability and predictive validity of behavioral tests. *Appl. Anim. Behav. Sci.* **127**, 51–65. doi: 10.1016/j.applanim.2010.08.007.

[16] Riemer S, Müller C, Virányi Z, Huber L, Range F. 2014. The predictive value of early behavioural assessments in pet dogs – a longitudinal study from neonates to adults. *PLoS ONE* **9**, e101237. doi: 10.1371/journal.pone.0101237.

[17] Kis A, Klausz B, Persa E, Miklósi Á, Gácsi M. 2014. Timing and presence of an attachment person affect sensitivity of aggression tests in shelter dogs. *Vet. Rec.* **174**, 196. doi: 10.1136/vr.101955.

[18] Serpell JA, Duffy DL. 2016. Aspects of juvenile and adolescent environment predict aggression and fear in 12-month-old guide dogs. *Front. Vet. Sci.* **3**. doi: 10.3389/fvets.2016.00049.

[19] Robinson LM, Thompson RS, Ha JC. 2016. Puppy temperament assessments predict breed and American Kennel Club group but not adult temperament. *J. Appl. Anim. Welf. Sci.* **19**, 101–114. doi: 10.1080/10888705.2015.1127765.

[20] Marder AR, Shabelansky A, Patronek GJ, Dowling-Guyer S, D'Arpino SS. 2013. Food-related aggression in shelter dogs: a comparison of behavior identified by a behavior evaluation in the shelter and owner reports after adoption. *Appl. Anim. Behav. Sci.* **148**, 150–156. doi: 10.1016/j.applanim.2013.07.007.

[21] Mohan-Gibbons H, Weiss E, Slater M. 2012. Preliminary investigation of food guarding behavior in shelter dogs in the United States. *Animals* **2**, 331–346. doi: 10.3390/ani2030331.

[22] Mornement KM, Coleman GJ, Toukhsati SR, Bennett PC. 2015. Evaluation of the predictive validity of the Behavioural Assessment for Re-homing K9's (B.A.R.K.) protocol and owner satisfaction with adopted dogs. *Appl. Anim. Behav. Sci.* **167**, 35–42. doi: 10.1016/j.applanim.2015.03.013.

[23] Poulsen AH, Lisle AT, Phillips CJC. 2010. An evaluation of a behaviour assessment to determine the suitability of shelter dogs for rehoming. *Vet. Med. Int.* **2010**, e523781. doi: 10.4061/2010/523781.

[24] Rayment DJ, Groef BD, Peters RA, Marston LC. 2015. Applied personality assessment in domestic dogs: limitations and caveats. *Appl. Anim. Behav. Sci.* **163**, 1–18. doi: 10.1016/j.applanim.2014.11.020.

[25] Hennessy MB. 2013. Using hypothalamic–pituitary–adrenal measures for assessing and reducing the stress of dogs in shelters: a review. *Appl. Anim. Behav. Sci.* **149**, 1–12. doi: 10.1016/j.applanim.2013.09.004.

[26] Protopopova A. 2016. Effects of sheltering on physiology, immune function, behavior, and the welfare of dogs. *Physiol. Behav.* **159**, 95–103. doi: 10.1016/j.physbeh.2016.03.020.

[27] Stephen JM, Ledger RA. 2005. An audit of behavioral indicators of poor welfare in kenneled dogs in the United Kingdom. *J. Appl. Anim. Welf. Sci.* **8**, 79–95. doi: 10.1207/s15327604jaws0802_1.

[28] Rooney NJ, Gaines SA, Bradshaw JWS. 2007. Behavioural and glucocorticoid responses of dogs (*Canis familiaris*) to kennelling: investigating mitigation of stress by prior habituation. *Physiol. Behav.* **92**, 847–854. doi: 10.1016/j.physbeh.2007.06.011.

[29] Patronek GJ, Bradley J. 2016. No better than flipping a coin: reconsidering canine behavior evaluations in animal shelters. *J. Vet. Behav.* **15**, 66–77. doi: 10.1016/j.jveb.2016.08.001.

[30] Protopopova A, Wynne CDL. 2014. Adopter-dog interactions at the shelter: behavioral and contextual predictors of adoption. *Appl. Anim. Behav. Sci.* **157**, 109–116. doi: 10.1016/j.applanim.2014.04.007.

[31] Martin JGA, Pirottay E, Petellez MB, Blumstein DT. 2017. Genetic basis of between-individual and within-individual variance of docility. *J. Evol. Biol.* **30**. doi: 10.1111/jeb.13048.

[32] Kruschke JK. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan.* Academic Press.

[33] McElreath R. 2015. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC Press.

[34] Voith VL et al. 2013. Comparison of visual and DNA breed identification of dogs and inter-observer reliability. *Am. J. Sociol.* **3**, 17–29. doi: 10.1080/10888700902956151.

[35] Goold C, Newberry RC. 2017. Aggressiveness as a latent personality trait of domestic dogs: testing local independence and measurement invariance. *bioRxiv*. doi: 10.1101/117440.

[36] Owczarczak-Garstecka SC, Burman OH. 2016. Can sleep and resting behaviours be used as indicators of welfare in shelter dogs (*Canis lupus familiaris*)? *PLoS ONE* **11**, e0163620. doi: https://doi.org/10.1371/journal.pone.0163620.

[37] R Development Core Team. 2016. *R: a language and environment for statistical computing*. Vienna, Austria.

[38] Tastle WJ, Wierman MJ. 2007. Consensus and dissention: a measure of ordinal dispersion. *Int. J. Approx. Reason.* **45**, 531–545. doi: 10.1016/j.ijar.2006.06.024.

[39] Ruedin D. 2016. *agrmt: calculate agreement or consensus in ordered rating scales*. R package version 1.40.4.

[40] Liddell TM, Kruschke JK. 2015. Analyzing ordinal data: support for a Bayesian approach. *SSRN*. doi: http://dx.doi.org/10.2139/ssrn.2692323.

[41] Foulley JL, Jaffrézic F. 2010. Modelling and estimating heterogeneous variances in threshold models for ordinal discrete data via Winbugs/Openbugs. *Comput. Methods Programs Biomed.* **97**, 19–27. doi: 10.1016/j.cmpb.2009.05.004.

[42] Kizilkaya K, Tempelman RJ. 2005. A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genet. Sel. Evol.* **37**, 31. doi: 10.1186/1297-9686-37-1-31.

[43] Nakagawa S, Schielzeth H. 2010. Repeatability for gaussian and non-gaussian data: a practical guide for biologists. *Biol. Rev.* **85**, 935–956. doi: 10.1111/j.1469-185X.2010.00141.x.

[44] Brommer JE. 2013. Variation in plasticity of personality traits implies that the ranking of personality measures changes between environmental contexts: calculating the cross-environmental correlation. *Behav. Ecol. Sociobiol.* **67**, 1709–1718. doi: 10.1007/s00265-013-1603-9.

[45] Stan Development Team. 2016a. *Stan Modeling Language Users Guide and Reference Manual*. Version 2.15.0.

[46] Lewandowski D, Kurowicka D, Joe H. 2009. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001. doi: 10.1016/j.jmva.2009.04.008.

[47] Watanabe S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594. url: http://www.jmlr.org/papers/v11/watanabe10a.html.

[48] Stan Development Team. 2016b. *RStan: the R interface to Stan*. R package version 2.14.1.

[49] Hiby EF, Rooney NJ, Bradshaw JWS. 2006. Behavioural and physiological responses of dogs entering re-homing kennels. *Physiol. Behav.* **89**, 385–391. doi: 10.1016/j.physbeh.2006.07.012.

[50] Svartberg K, Forkman B. 2002. Personality traits in the domestic dog (*Canis familiaris*). *Appl. Anim. Behav. Sci.* **79**, 133–155. doi: 10.1016/S0168-1591(02)00121-1.

[51] Hsu Y, Serpell JA. 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *J. Am. Vet. Med. Assoc.* **223**, 1293–1300. doi: 10.2460/javma.2003.223.1293.

[52] Jones AC, Gosling SD. 2005. Temperament and personality in dogs (*Canis familiaris*): a review and evaluation of past research. *Appl. Anim. Behav. Sci.* **95**, 1–53. doi: 10.1016/j.applanim.2005.04.008.

[53] Riemer S, Müller C, Virányi Z, Huber L, Range F. 2016. Individual and group level trajectories of behavioural development in Border collies. *Appl. Anim. Behav. Sci.* **180**, 78–86. doi: 10.1016/j.applanim.2016.04.021.

[54] Cramer AOJ, Borkulo CD van, Giltay EJ, Maas HLJ van der, Kendler KS, Scheffer M, Borsboom D. 2016. Major Depression as a complex dynamic system. *PLoS ONE* **11**, e0167490. doi: 10.1371/journal.pone.0167490.

[55] Wichers M, Groot PC, Psychosystems, ESM Group, EWS Group. 2016. Critical slowing down as a personalized early warning signal for depression. *Psychother. Psychosom.* **85**, 114–116. doi: 10.1159/000441458.

[56] David JT, Cervantes MC, Trosky KA, Salinas JA, Delville Y. 2004. A neural network underlying individual differences in emotion and aggression in male golden hamsters. *Neuroscience* **126**, 567–578. doi: 10.1016/j.neuroscience.2004.04.031.

[57] Betini GS, Norris DR. 2012. The relationship between personality and plasticity in tree swallow aggression and the consequences for reproductive success. *Anim. Behav.* **83**, 137–143. doi: 10.1016/j.anbehav.2011.10.018.

[58] Dewitt TJ, Sih A, Wilson DS. 1998. Costs and limits of phenotypic plasticity. *Trends Ecol. Evol.* **13**, 77–81. doi: 10.1016/S0169-5347(97)01274-3.

[59] Highcock L, Carter AJ. 2014. Intraindividual variability of boldness is repeatable across contexts in a wild lizard. *PLoS ONE* **9**, e95179. doi: 10.1371/journal.pone.0095179.

[60] Araya-Ajoy YG, Dingemanse NJ. 2017. Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *J. Anim. Ecol.* **86**, 227–238. doi: 10.1111/1365-2656.12621.

[61] Dykiert D, Der G, Starr JM, Deary IJ. 2012. Age differences in intra-individual variability in simple and choice reaction time: systematic review and meta-analysis. *PLOS ONE* **7**, e45759. doi: 10.1371/journal.pone.0045759.