Predicting shelter dog behaviour after adoption using a longitudinal behavioural assessment

Conor Goold[1,2] and Ruth C. Newberry[2]

[1]Faculty of Biological Sciences, University of Leeds, UK, LS2 9JT
[2]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway

Abstract. Predicting the behaviour of shelter dogs after adoption is an important, but difficult, endeavour. Despite three decades of research on behaviour assessments in shelters, evidence supporting their use is mixed. In part, this has been attibuted to the use of behaviour test batteries that fail to represent the typical behaviour of shelter dogs across time. This study provides the first analysis of the ability for a longitudinal, observational behaviour assessment to predict behaviour post-adoption. We analyse shelter behavioural observations ($> 60,000$ records in total) and post-adoption behavioural reports (using telephone surveys) across eight contexts from 241 dogs. A novel joint hierarchical Bayesian mixture model is used to i) compare shelter and post-adoption behaviour across dogs and across contexts, ii) model patterns of missing data, and iii) estimate measurement error in shelter and post-adoption behavioural records. Dog behaviour at the shelter correlated positively with behaviour post-adoption within contexts (0.38; 95% highest density interval: [0.20, 0.55]), and how dogs behaved within contexts explained more of the behavioural variance than behaviour between dogs or between contexts alone. We found few differences in individual-level probabilities of different behaviours post-adoption compared to at the shelter. However, part of this good predictive ability arose from accurate representation of uncertainty in individual-level behavioural predictions, not accounted for when inspecting the raw data. Measurement error was higher post-adoption than at the shelter. Longitudinal assessments can make reasonably accurate predictions about post-adoption behaviour, but uncertainty, between-dog heterogeneity and measurement error should be appropriately accounted for.

Contents

1 Introduction

Accurately predicting the future behaviour of domestic dogs (*Canis familiaris*) is a priority for professional organisations, dog breeders, and dog owners alike. Animal shelters have the difficult task of assessing how dogs will behave in a variety of circumstances from limited behavioural information, and making decisions about the suitability of those dogs for placement into new homes. To do so, animal shelters frequently employ standardised behaviour tests that evaluate specific behavioural or *personality* traits through reconstructions of situations relevant to life outside the shelter environment (van der Borg et al. 1991; Marston and Bennett 2003; Mornement et al. 2010; Taylor and Mills 2006; Rayment et al. 2015).

For example, to assess a dog's level of aggressiveness around food (i.e. food guarding), shelter staff may present a dog with a bowl of food and record the dog's response to a plastic hand approaching, touching or trying to remove the food bowl (Mohan-Gibbons et al. 2012; Mohan-Gibbons et al. 2018; Marder et al. 2013). Standardised tests are usually conducted once, soon after arrival at the shelter, yielding an overall score used to help determine the dog's suitability for adoption.

Despite three decades of research on the development, implementation and predictive validity of shelter dog behavioural assessments, the evidence supporting their use is mixed. The reliability and validity of many behavioural assessments used by shelters have not been ascertained (Taylor and Mills 2006; Mornement et al. 2009; Mornement et al. 2010; Mornement et al. 2014). The feasibility of carrying out behavioural assessments with high reliability and standardisation in the time-constrained shelter environment has also been questioned, with assessments often taking at least an hour per dog (van der Borg et al. 1991; Mornement et al. 2009; although see Poulsen et al. (2010) for a shorter evaluation). In the last ten years, criticisms have been levelled against the ability for behavioural assessments to predict dog behaviour after adoption with reasonable accuracy. A high number of false positives have been recorded for food guarding behaviour (i.e. displaying food aggression during shelter evaluations but not in the new home; see Mohan-Gibbons et al. 2012; Marder et al. 2013) and has led to calls for abandoning standardised food tests after return rates to shelters did not increase when the tests had been terminated (Mohan-Gibbons et al. 2018). Other types of aggressive and non-aggressive behaviours might be similarly difficult to predict post-adoption (Christensen et al. 2007; Mornement et al. 2015; although see Bollen and Horowitz (2008)).

The reasons why dog behaviour is difficult to predict post-adoption are varied, and ways to improve predictions are not entirely clear. Widespread advice to improve the reliability and validity of shelter dog assessments are hampered by a lack of clarity of what those terms mean (Patronek et al. 2019; Rayment et al. 2015), which is also a problem in the wider scientific community (e.g. see Borsboom et al. 2004; Borsboom et al. 2009; Maul et al. 2016). Patronek and Bradley (2016) demonstrate that even with high sensitivity and specificity, the probability of a dog showing aggression in the new home after a positive test in the shelter is unlikely to be higher than 50%, and is probably closer to $\sim 30\%$ (see also Patronek et al. 2019). Tests that have been reported to predict behaviour successfully post-adoption (e.g. Valsecchi et al. 2011; Poulsen et al. 2010; Bollen and Horowitz 2008) support their claims using statistical significance of linear associations (e.g. significant correlations or regression coefficients) between shelter and post-adoption behaviour (i.e. predictive 'validity'; Patronek et al. 2019). In constrast, studies arguing against the efficacy of behavioural tests do so with evidence of low predictive 'ability', i.e. estimates of true/false positive and negative rates (Patronek et al. 2019). As an alternative, a number of authors and organisations emphasise the collection of daily behavioural observations from dogs in shelters, in addition to other information (e.g. foster reports, pre-surrending interviews), to formulate a holistic profile of each dog's behaviour and welfare needs used to inform adoption (Patronek et al. 2019; ASPCA 2018; Rayment et al. 2015; Mornement et al. 2015; Clay et al. 2020). Longitudinal approaches refrain from crudely labelling individual dogs as, for example, aggressive or non-aggressive from single behavioural tests. However, research is still scarce on how best to implement this approach and summarise the swathes of information observational methods generate on each dog to be of practical use for shelters.

In previous studies, we have reported on the behaviour of dogs in shelters from a longitudinal and observational assessment methodology implemented in one UK shelter (Goold and Newberry 2017b; Goold and Newberry 2017a). The assessment relies on the sponataneous collection of behavioural observations from everyday occurrences (e.g. walking past the dog's kennel, putting a food bowl into the kennel, clipping on the lead or touching the collar) to cumulatively determine how best to maintain the dog's welfare and place into a new home. Goold and Newberry (2017a) used the framework of behavioural reaction norms (Dingemanse et al. 2010; Cleasby et al. 2015) to partition individual variation in over 3,000 dogs' behaviours ($\sim$ 20,000 behavioural reports in total) around unknown people into personality (i.e. average behaviour), plasticity (i.e. behavioural change over time at the shelter) and predictability (i.e. residual variance). Accounting for all three components improved the out-of-sample predictive accuracy of the statistical models, due to large uncertainty imposed by differences in behaviour between dogs and variable amounts of behavioural information available on each dog.

While accounting for individual variation in behaviour across time at the shelter is important, making predictions about dog behaviour post-adoption from longitudinal information brings its own challenges. Deciding on what statistical moments (e.g. cumulative mass, central tendency, scale) of the longitudinal shelter data should predict post-adoption behaviour is not obvious. Diagnostic statistics, such as false positives or false negatives, lose their clarity because longitudinal assessments acknowledge that dog behaviour can change depending on time and context (Goold and Newberry 2017a), and thus understanding behavioural patterns (e.g. how dogs behave on average) is more important than occurrences of single behaviours. The realities of collecting observational data in busy shelter environments also means that data may present significant patterns of non-ignorable missingness, substantially unbalanced repeat measures, and could be subject to more measurement error than standardised assessments. Altogether, determining the predictive value of 'on-going' shelter assessments will require careful analysis of longitudinal data using statistical models that appropriately account for the data-generating processes.

The goal of the present study was to assess the correspondance between behaviour shown at the shelter, collected using the longitudinal observation methodology, and the behaviour reported during telephone interviews post-adoption. Our secondary goal is more methodological, as we present a novel application of joint Bayesian hierarchical modelling to make

Demographic variable	**Mean (SD)/n**
Number of total observations per dog	140.7 (224.3)
Days spent at the shelter	32.1 (41.4)
Estimated age at departure (years)	3.6 (2.8)
Weight (kg; last recorded at shelter)	16.5 (9.6)
Source: gift/stray/return	164/51/26
Rehoming centre: London/Old Windsor/Brands Hatch	71/93/77
Sex: female/males	102/139
Neutered: before arrival/at shelter/not/missing	84/145/11/1

Table 1: Dog demographic variables collected at the shelter.

predictions about dog behaviour post-adoption. Importantly, our approach accounts for i) individual variation in both average behaviour (i.e. personality) and behavioural change (i.e. plasticity; due to only two post-adoption interviews, predictability was not estimated) across shelter and post-adoption time periods, ii) missing data and their potential correlates, and iii) measurement error in the behavioural reports. We report on predictive validity by estimating correlations between personality and plasticity between shelter and post-adoption time periods. Predictive ability is examined by comparing model predictions for each dog's average behaviour in the shelter and post-adoption to determine whether dogs exhibited credible changes in their behaviour. We present specific examples of predicting individual-level behaviour in our results and discussion.

2 Materials & Methods

2.1 Subjects

Behavioural data were gathered retrospectively from new owners on 265 dogs adopted from Battersea Dogs and Cats Home, United Kingdom between May 2016 and May 2017. Details of the shelter environment can be found in Goold and Newberry (2017b) and Goold and Newberry (2017a). Of the 265 dogs, only 241 dogs could be matched with behavioural and demograhpic data from the shelter's database. Demographic information about these dogs is presented in Table 1. Demographic information of the new owners was collected but not analysed here due to inconsistencies in the responses.

2.2 Data collection

2.2.1 Shelter behaviour

As previously described by Goold and Newberry (2017b) and Goold and Newberry (2017a), shelter employees observed dogs in a variety of naturally-occurring contexts (Table 2) and recorded each dog's behaviour using a context-specific list of mutually-exclusive behavioural codes. The contexts analysed here were eight core 'on-site' contexts. The *Interactions with dogs* context was a combination of the original *Interactions with female* and *male dogs* contexts, respectively, because the post-adoption data did not distinguish between interactions with male and female dogs (as adopters may not have been fully aware of the sex of other dogs that their dog met). Observations were made as often as possible given the frequency the situation occurred and the time constraints on entering data for shelter staff. However, we consider any day the dog did not receive an observation in a context as a missing observation so that patterns of missingness could be modelled (see Table 2 for amount of missing daily observations). In total, there were $61,952$ (missing and complete) observations in the shelter dataset.

There were between ten and sixteen possible behavioural codes depending on the context, arranged on a scale of perceived ease of adoption or desirability to adopters (see Supplementary Materials for the full list of behavioural codes). The behavioural codes were further categorised by the shelter into green, amber and red categories: green codes indicated that the dog's behaviour was suitable for immediate adoption and/or would be easy for all owners to manage once adopted (e.g. the dog was friendly or excited when meeting new people), amber codes indicated that some training or management might be needed to manage or improve behaviour (e.g. the dog was nervous when meeting new people and slow to meet with them), and red codes that the dog needed an individualised training program and improvement in behaviour to facilitate adoption or would be the most difficult for adopters to manage (e.g. the dog showed aggression when meeting unfamiliar people). A dog's suitability for adoption was decided based on behaviour across all contexts and days after arrival. Dogs were not adopted to the public if they were considered to pose a serious and highly probable risk towards people or other dogs.

The behavioural scores were analysed using the green, amber and red ordinal scale, rather than using the individual behavioural codes of each ethogram. This was chosen: i) to place the behaviour records across contexts on a comparable scale, allowing all the data to be analysed within the same statistical model; ii) to capture the main variation in the data given that some individual codes were seldom used; iii) because previous analysis indicated higher consensus among shelter staff when rating behaviour using the green, amber and red codes than the individual codes (see Goold and Newberry 2017a); and iv) it was more practically relevant because the shelter was interested to find out if their assessments at the shelter differed greatly from a dog's behaviour post-adoption (i.e. there was a change from

Context	Description	Median (IQR)	% Missing (SD)
Handling (HND)	Informal handling by people (e.g. stroking non-sensitive areas, touching or holding the collar, fitting a harness or lead).	13 (14)	40 (20)
In kennel (KNL)	Inside the kennel	14 (14)	37 (21)
Outside the kennel (OKNL)	Outside the kennel	13 (14)	42 (19)
Interactions with familiar people (FPL)	Interacts with familiar people (interacted with at least once before) outside the kennel who approach, make eye contact, speak to or attempt to make physical contact with the dog.	13 (14)	44 (18)
Interactions with unfamiliar people (UFPL)	Interacts with familiar people (never interacted with before) outside the kennel who approach, make eye contact, speak to or attempt to make physical contact with the dog.	6 (5)	75 (15)
Eating food (FOOD)	Eating food from a container, toy or while being hand-fed	14 (13)	39 (20)
Interactions with toys (TOYS)	Interacting with dog toys	8 (13)	58 (25)
Interactions with dogs (DOGS)	Meeting dogs outside the kennel during structured interactions and/or sponataneous meetings	5 (4)	76 (12)

Table 2: Observational contexts at the shelter with corresponding median and interquartile range (IQR; due to skewness) of the number of records per dog, and average percentage (standard deviation) of missing daily records per dog.

green to red codes), but not necessarily if there was a change between codes within the same colour category. When more than one record was made of the same dog in the same context on the same day, the most ‘severe’ code assigned was retained for analysis.

2.2.2 Post-adoption behaviour

On the point of adoption, adopters were asked to participate in a study evaluating the predictive accuracy of the shelter’s behavioural assessment and were given full details of the procedure (consent form provided in the Supplementary Materials). Each consenting adopter received two phone calls from a canine behaviourist at the shelter who surveyed them about their dog’s behaviour using a standardised set of questions (provided in the Supplementary Materials). The majority of phone calls were made by one behaviourist although no records were kept on which behaviourist made the calls. The study planned to record behaviour using telephone interviews at approximately 2-3 weeks and 5-6 weeks after adoption, but the real days after adoption were more variable. The average number of days after adoption for the first phone call was 19.8 (sd = 4.9; range = 7 to 51), and 32.7 days (sd = 11.1; range = 14 to 72) for the second, with an average of 20.7 (sd = 5.6; range = 4 to 42) days between first and second surveys for those dogs with two completed surveys. Only 150 dogs (62%) had two completed surveys. For one dog that had been returned after their post-adoption surveys, and was subsequently rehomed and received new surveys from their second owners, we randomly chose their first set of surveys to be included in the analysis.

The first questions gathered information about the adopters' dog ownership experience and the post-adoption environment. The remaining questions (Table 3) enquired about how the dog reacted in situations comparable to the eight behavioural observation contexts in the shelter assessment (Table 2). The questions were used as a guide during the post-adoption interviews, with the behaviourists providing more detailed descriptions as needed. One additional context ('*How has the dog behaved when with any other resource?*') had too few records in either the post-adoption reports or at the shelter to be included in the analysis. The *In kennel* and *Out of kennel* contexts at the shelter were transformed to *In house* and *Out of house*. The questions were open-ended, with the behaviourist encouraging adopters to describe the dog's behaviour in each situation rather than label the behaviour. Subsequently, the behaviourist chose the behavioural code for each context that best described the behaviour. Adopters were allowed to respond 'No opinion'. Phone calls were not recorded, precluding assessment of the reliability of behaviour coding during phone calls. Data were handled anonymously by the authors, who received only the dogs' identification numbers to enable matching of the post-adoption reports with the shelter records.

2.3 Statistical analysis

2.3.1 Theoretical approach

The green, amber and red codes from shelter and post-adoption time points were analysed using a joint or 'shared parameter' (e.g. see Vonesh et al. 2006; Tseng et al. 2016) Bayesian hierarchical model, which accounted for two data complexities. First, we treated the missing data present in the shelter records (Table 2) and post-adoption surveys (Table 3) as non-ignorable (i.e. not random) because they likely depended on other variables. For example, the number of people unfamiliar to a dog decreases with time spent at the shelter or time after adoption, leading to potentially more missing records in the *Interactions with unfamiliar people* context. A dog's overall behaviour could also impact missing values. For example, dogs who have more green codes overall and do not change their behaviour may not receive as many records due to a lack of priority over dogs who show more worrying behaviour. The participation or attrition rate of adopters in the telephone surveys may also have depended on their dog's behaviour post-adoption.

Context (abbreviation)	**Survey question**	**n (%) 2 surveys**
Handling (HND)	How has the dog behaved whilst being handled or restrained (informal)?	149 (61.8)
In house (HOUSE)	How has the dog behaved in the house?	149 (61.8)
Outside of house (OTSD)	How has the dog behaved outside on walks?	147 (61.0)
Interactions with familiar people (FPL)	How has the dog behaved when meeting familiar people?	144 (59.8)
Interactions with unfamiliar people (UFPL)	How has the dog behaved when meeting unfamiliar people	149 (61.8)
Eating food (FOOD)	How has the dog behaved with their food?	150 (62.2)
Interactions with toys (TOYS)	How has the dog behaved with toys?	149 (61.8)
Interactions with dogs (DOGS)	How has the dog behaved when meeting another dog for the first time?	144 (59.8)

Table 3: Post-adoption survey behaviour questions and response rates (number and percentage of owners responding to the question over two surveys).

A second complexity was the high proportion ($\sim 90\%$) of green codes, which was found in previous analysis of the same shelter assessment (Goold and Newberry 2017a). We hypothesised here that some proportion of the green codes could have been recorded due to other processes, such as staff members not observing the dog's behaviour directly, or forgetting how a dog behaved, but inputting a green code anyway. Adopters may have also described their dog's behaviour in terms consistent with green codes (i.e. not reporting any problems) during phone calls when the dog's behaviour might be more consistent with amber or red code behaviour. Thus, if not missing, some green codes were potentially 'inflated', leading to a greater probability mass for green codes beyond that explained by the data generating processes of the behavioural scale alone. Similar assumptions are applied to count data with a high number of zero values (zero inflation; Lambert 1992), and have also been used to understand a high probability mass for 'I don't know' responses (Kelley and Anderson 2008) and 'face-saving' or 'Neither agree nor disagree' responses (Bagozzi and Mukherjee 2012) on ordinal survey scales.

To account for the above complexities, we specified a custom mixture model for the probability of different codes ($c = \text{missing}, \text{green}, \text{amber}, \text{red} = 0, 1, 2, 3$) for case i (either the day at the shelter or the day after adoption), dog j and context k:

$$p(y^c_{ijk}) = \begin{cases} \psi_{ijk} & \text{if } y_{ijk} = 0 \\ (1 - \psi_{ijk})\big[\kappa + (1 - \kappa)\pi^c_{ijk}\big] & \text{if } y_{ijk} = 1 \\ (1 - \psi_{ijk})(1 - \kappa)\pi^c_{ijk} & \text{if } 1 < y_{ijk} \leq 3 \end{cases} \tag{1}$$

The parameter ψ_{ijk} is a 'hurdle' probability of missing data for each dog and context (see Kubinec 2019 for a similar approach to handling missing data), which is modelled as a Bernoulli trial from the data:

$$y^{c}_{ijk}\big|_{c=0} \sim \text{Bernoulli}(\psi_{ijk}) \tag{2}$$

The complement, $1-\psi_{ijk}$, is the probability of either a staff member choosing to input an observation in the shelter or an adopter participating in the telephone survey, respectively. The κ parameter is a mixing term that describes the probability of a green code being drawn from the inflation component. Consequently, $1-\kappa$ is the probability of a code being non-inflated. A non-missing and non-inflated code of cateogory c (and for a specific case, dog and context) occurs with probability $(1-\psi_{ijk})(1-\kappa)\pi^{c}_{ijk}$, where π^{c}_{ijk} is defined using a cumulative ordinal probit model:

$$y^{c}_{ijk}\big|_{c=\{1,2,3\},1-\kappa} \sim \text{Categorical}(\pi^{c}_{ijk}) \tag{3}$$

$$\pi^{c}_{ijk} = \Phi\Big(\frac{\theta_c - \mu_{ijk}}{\sigma}\Big) - \Phi\Big(\frac{\theta_{c-1} - \mu_{ijk}}{\sigma}\Big) \tag{4}$$

Equation 3 describes the non-missing and non-inflated code as categorically distributed with probability π^{c}_{ijk}, which is defined in equation 4 as the cumulative area under a latent standard normal distribution (where Φ is the standard normal cumulative distribution function) between threshold parameters θ_c and θ_{c-1} ($\boldsymbol{\theta} = \theta_0, \theta_1, \theta_2, \theta_3$). The probability of the first and last threshold parameters (i.e. $\Phi(\theta_0)$, $\Phi(\theta_3)$) were set to 0 and 1, respectively. To estimate the mean (μ_{ijk}) and standard deviation (σ) on the scale of the ordinal data, we fixed $\theta_1 = 1.5$ and $\theta_2 = 2.5$ (see Kruschke 2014).

2.3.2 Joint modelling

The data from the shelter and post-adoption time periods were each analysed by the model described above, which we describe as two distinct 'sub-models'. Each sub-model included dog-, context-, and dog $\times$ context-varying intercept and slope (for the days after arrival or days after adoption, respectively) parameters. The joint aspect of our model derives from correlations between these parameters across the shelter and post-adoption sub-models. For the ordinal probit regressions (equation 4), we included random intercepts and slope terms for dogs, contexts and the interaction between dogs and contexts ($1,928$ unique dog $\times$ contexts combinations). Dog- and context-varying effects capture additive variation due to dogs and contexts, respectively, while the dog $\times$ context interaction describes the non-additive, unique variation attributed to specific dogs in specific contexts. For the missing data regressions, we included only random intercepts. For each random effect term (dogs, contexts, and dogs $\times$ contexts), this led to a 6 x 6 covariance matrix capturing the relationship between random intercepts and slopes for the shelter and post-adoption behaviours.

2.3.3 Predictor variables

For the shelter sub-model, we included days after arrival as an observation-level predictor and the variables from Table 1, save the total number of observations (which was nearly perfectly collinear with length of stay) and rehoming centre (which was not of interest), as dog-level predictors on both the probability of missing data (ψ_{ijk} in equation 2 using a logit link function) and the predicted mean of the latent behavioural scale (μ_{ijk} in equation 4 using the identity link function). For the post-adoption sub-model, we predicted the probability of missing data and the latent behavioural scale scores using days after adoption as an observation-level predictor and the same shelter demographic variables as in the shelter sub-model as retrospective dog-level predictors, as well as the total number of surveys the dog had as an additional dog-level predictor. Sum-to-zero coding was used for categorical predictors. Dog-level metric predictors were mean-centered and standardised by 1 standard deviation. Days after arrival to the shelter was centered around the average length of stay at the shelter (32.1 days; Table 1) and scaled by 1 standard deviation (73 days). Days after adoption was centered around the average of each dog's mean number of days after adoption (26.3 days) and scaled by its standard deviation ($\sim$ 12 days). One dog had missing neuter status data, two dogs were missing weight data and two dogs were missing age at departure data. Due to the small amount, we imputed the missing neuter status data point with the most frequent category (neutered on site; Table 1), and mean imputed the missing weight and age data points.

2.3.4 Estimation & inference

All data cleaning and post-processing was conducted in R version 3.6.1 (R Core Team 2019). We fit our model using Bayesian estimation in the Stan programming language (Stan Development Team 2018) using the terminal interface CmdStan version 2.23.0 (Stan Development Team et al. 2018). Stan employs Hamiltonion Monte Carlo, a type of Markov chain Monte Carlo (MCMC) sampling algorithm, to sample from the posterior distribution. After ensuring adequate mixing of multiple chains from the model ran on smaller sub-sets of the data, we took 10,000 iterations from the posterior distribution using 2 MCMC chains (1000 warmup and 5000 sampling). We summarise parameters by their means and 95% highest density intervals (HDIs). For the ordinal data, we make predictions using both the latent metric scale (parameters denoted by β) and the corresponding posterior probabilities of each ordinal category (probabilities denoted by π). When evaluating population-level parameters, we highlight discrepancies, where appropriate, between the estimates for average dogs and contexts, and estimates that incorporate uncertainty due to variation across dogs and contexts. The former represent estimates made at the mean of the random effects, while the latter are made by computing predictions marginal of the random-effect distributions and are particularly important for making accurate future predictions in hierarchical models (e.g. IntHout et al. 2016; Wang and Lee 2019).

2.4 Ethical statement

Approval for the processing of personal data was gained from the Norwegian Social Science Data Services (approval number 47080). Names and addresses of the participating new owners and their dogs were anonymised before data were passed to the authors.

2.5 Data & code accessibility

We provide complete mathematical details of the above model, supplementary files, the data, the Stan code, and the R code to reproduce the results reported here at `https://github.com/ConorGoold/GooldNewberry2020-lba`. Due to the non-standard statistical model, we also provide R and Stan code to simulate data and fit the model to recover parameters values at the same repository.

3 Results

3.1 Average behaviour

Green codes accounted for more than 95% of the observations in the raw shelter and post-adoption datasets, and for an average dog in an average context (i.e. at the mean of the random effects, and where all dog-level predictors equal 0), there were no credible differences between the probability of green, amber and red codes at the shelter (at approximately 30 days after arrival) and post-adoption (at approximately 30 days post-adoption; probability differences: $\pi_{\text{green}} = 0.00$, 95% HDI = [-0.05, 0.06]; $\pi_{\text{amber}} = 0.00$, 95% HDI: [-0.06, 0.05]; $\pi_{\text{red}} = 0.00$, 95% HDI: [0.00, 0.00]). Similarly, for an average dog in an average context, behaviour tended to improve with every additional week at the shelter ($\beta = -0.08$; 95% HDI: [-0.14, -0.02]), although due to ceiling effects the practical increase in green codes was minimal (from around 98% on arrival day to 99% likely on day 30). There was no clear relationship of days after adoption on post-adoption behaviour ($\beta = 0.01$; 95% HDI: [-0.21, 0.24]). A dog's average probability of missing data at the shelter was $\pi_{\text{missing}} = 0.54$ (95% HDI: [0.37, 0.73]), and the odds of missing codes increased with every additional week after arrival (odds: 1.04; 95% HDI: [1.03, 1.05]). The odds of missing data post-adoption increased sharply with every additional week after adoption (odds: 11.62; 95% HDI: [8.90, 14.00]), although the baseline probabilities of missing data at 30 days post-adoption between dogs with one survey were $\pi_{missing} = 0.45$ (95% HDI: [0.26, 0.65]) compared to essentially 0 probability of missing data for dogs with two surveys, highlighting that within-survey missing data was highly infrequent.

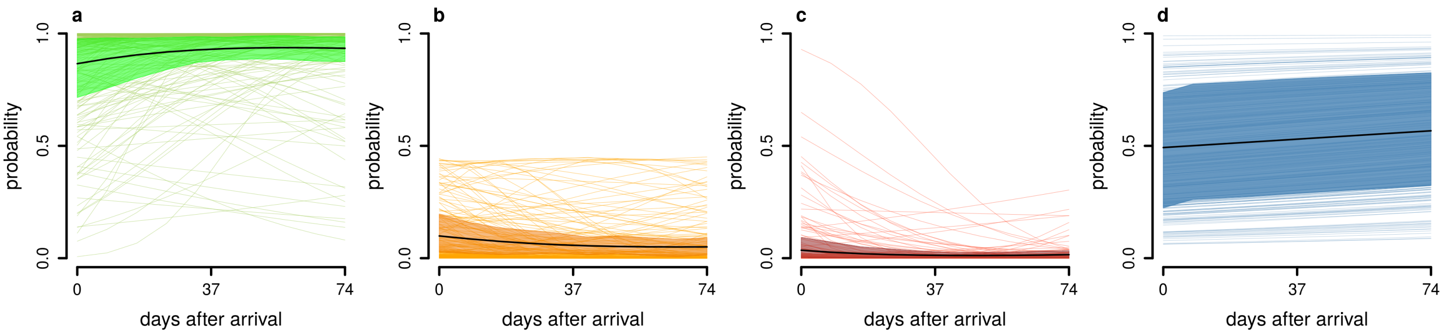

Figure 1: Probability of green (a), amber (b), red (c) and missing (d) codes across days after arrival to the shelter (average length of stay + 1 standard deviation) accounting for variation across dogs, contexts and dog × context behaviour. Thick black lines show posterior mean estimates, coloured bands show the 95% highest density interval of the mean, and thin coloured lines (n = 241) show one random sample from one randomly-chosen context from each dog's posterior distribution.

The probability of green-code inflation at the shelter was $\kappa = 0.04$ (95% HDI: [0.00, 0.09]) at the shelter, and $\kappa = 0.20$ (95% HDI: [0.03, 0.36]) post-adoption.

3.2 Variation across contexts

Incorporating variation across dogs and contexts demonstrated substaintially more uncertainty in the model estimates both at the shelter (Figure 1) and post-adoption (Figure 2). At the shelter, the *Interactions with dogs* context had the highest probability of red codes ($\pi_{red} = 0.04$; 95% HDI: [0.02, 0.06]; Figure S1a & Table S1), and dogs changed their behaviour the most in the *Handling* context ($\beta = -0.20$ 95% HDI: [-0.28, -0.14]; Figure S1b & Table S1). Post-adoption, the *Eating food* context had the highest probability of red codes ($\pi_{red} = 0.01$; 95% HDI: [0.00, 0.05]; Figure S2a & Table S2), although there were no strong differences between contexts in the average amount of behavioural change across days after adoption (Figure S2b & Table S2). Missing codes were particularly prominent at the shelter for the *Interactions with unfamiliar people* context ($\pi_{missing} = 0.81$; 95% HDI: [0.74, 0.89]).

3.3 Interaction between dogs and contexts

Both at the shelter and post-adoption, behaviour of dogs within contexts (the dog × context-varying intercepts) varied more than between dogs across all contexts or between contexts across all dogs. This is illustrated by larger repeatabilities (i.e. proportion of total variance explained; see Supplementary Materials for calculation) for the dogs × context-varying intercept parameters compared to the additive effects of dogs or contexts alone (Figure 3a; Table S3). Repeatability post-adoption was approximately 20% higher (but more uncertain) post-adoption for dogs within contexts than at the shelter ($R_{diff} = -0.23$; 95% HDI: [-0.41,

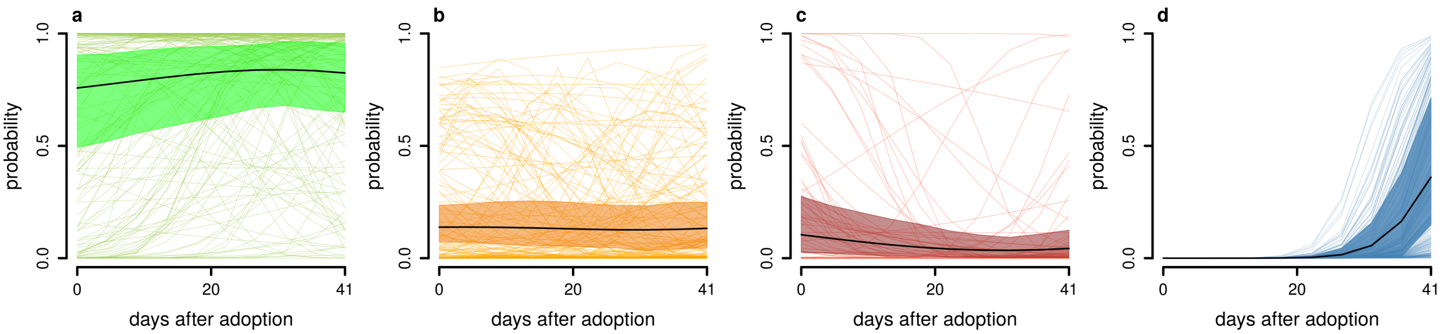

Figure 2: Probability of green (a), amber (b), red (c) and missing (d) codes across days after adoption (average days after adoption interview + 1 standard deviation) accounting for variation across dogs, contexts and dog × context behaviour. Thick black lines show posterior mean estimates, coloured bands show the 95% highest density interval of the mean, and thin coloured lines (n = 241) show one random sample from one randomly-chosen context from each dog's posterior distribution.

-0.05]) because the standard deviation of the dog × context-varying intercepts was 1.40 (95% HDI: [1.10, 1.72]) times higher post-adoption than at the shelter. The standard deviation of the dog × context slope parameters was considerably larger at the shelter (describing behaviour over days after arrival) than post-adoption (describing behaviour over days after adoption), but had substantial uncertainty (ratio of standard deviations: 45.54; 95% HDI: [1.61, 96.43]).

3.4 Random effect correlations

The dog × context-varying intercept parameters at the shelter and post-adoption had a moderate postive correlation ($\rho = 0.38$; 95% HDI: [0.20, 0.55]; Figure 3b), whereas the correlations between dog-varying and context-varying intercepts were largely positive but their 95% HDIs spanned zero (Table S3). Correlations between the random-slope parameters at the shelter and post-adoption included zero. At the shelter, the dog × context intercept and slope parameters had a positive correlation ($\rho = 0.26$; 95% HDI: [0.02, 0.47]), meaning dogs with higher values on the latent behavioural scale (higher changes of red codes) changed their behaviour less over time or, in some cases, were more likely to exhibit amber and red codes over time. Dogs with more positive slope parameters across contexts at the shelter had slightly higher chances of missing data at the shelter ($\rho = 0.28$; 95% HDI: [0.01, 0.54]). Dogs with fewer missing data on average post-adoption also had larger intercept parameters ($\rho = -0.38$; 95% HDI: [-0.71, -0.04]) and more positive slopes ($\rho = -0.36$; 95% HDI: [-0.69, -0.02]) at the shelter, suggesting dogs with overall higher chances of amber and red codes at the shelter had better response rates in the post-adoption phone calls. We note that average post-adoption behaviour had a mostly positive correlation with the amount of missing data post-adoption, although its 95% HDI was wide and just included zero ($\rho = 0.25$; 95% HDI:

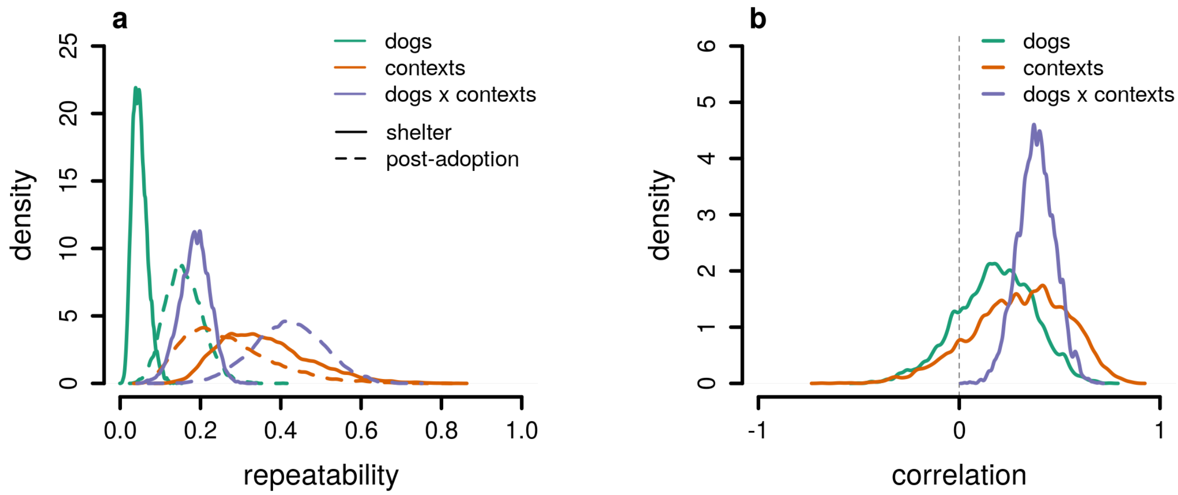

Figure 3: Proportions of total behavioural variance explained (repeatability) by dog-, context- and dog × context-varying intercepts for shelter and post-adoption time periods (a), and correlations for dog-, context- and dog × context-varying intercepts between shelter and post-adoption time periods (b). Each parameter is evaluated at approximately 30 days after arrival to the shelter and 30 days after adoption.

[-0.07, 0.53]).

3.5 Individual-level prediction of post-adoption behaviour

Behaviour was not static across shelter and post-adoption time points. A median of 48.50 (inter-quartile range: 32.75) and 39 (inter-quartile range: 66.75) dogs who received at least one amber and red code, respectively, within contexts at the shelter did not show amber or red codes post-adoption for the same contexts. Similarly, a median of 29.50 (inter-quartile range: 56.75) and 13 (inter-quartile range: 65.25) dogs who did not show red codes at the shelter within contexts, received red codes in the post-adoption reports (see Supplementary Materials for further diagnostics). However, these diagnostic values only provide a crude estimate of behavioural differences between the shelter and post-adoption time periods, ignoring unbalanced repeat measures, measurement error, missing data and dogs' underlying probabilities of showing different codes.

In contrast, we compared the posterior probabilities of green, amber and red codes on each dog's mean day after arrival to the shelter, with the same codes' predicted probabilities on the day of each dog's first and second telephone survey (where dogs with only the first survey had their second survey predictions estimated by the model). We summarised the resulting differences (δ) in code probabilties by counting the number and proportion of dogs whose 95% HDIs of the differences did not contain zero (Figure 4). In general, the number/proportion of dogs who had significantly different probabilities of green, amber or

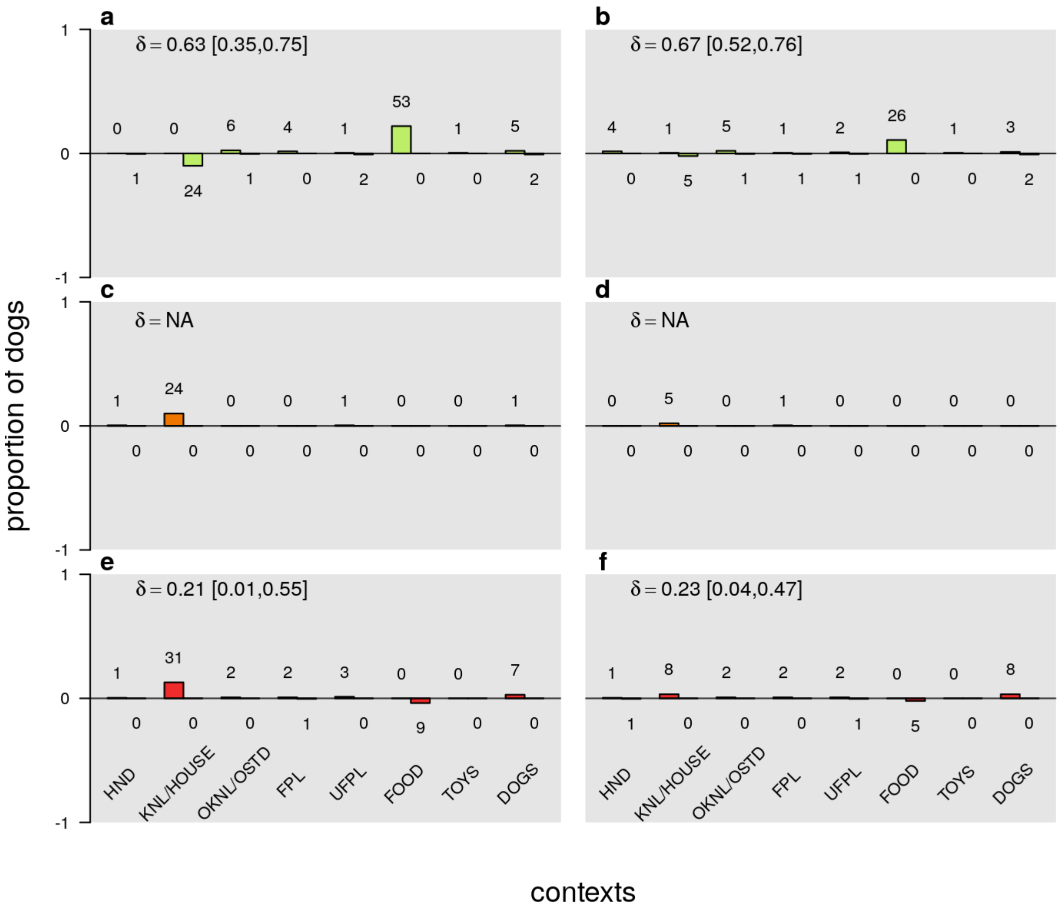

Figure 4: The proportion (bars) and number of dogs (text above each bar) who had significantly different probabilities of green (a-b), amber (c-d), and red (e-f) codes within contexts at the shelter compared to behaviour reported at the first (a, c, e) and second (b, d, f) post-adoption phone calls. Positive bars illustrate the proportion of dogs with increased code probabilities at the shelter compared to post-adoption; negative bars illustrate the proportion of dogs with decreased probabilities at the shelter compared to post-adoption. A dog was included in the count if the 95% highest density interval of the differences in code probabilities for a dog surpassed zero. The mean ($\delta = [0, 1]$) and 95% highest density interval absolute probability difference is shown above each plot, except for panels c and d where there were too few dogs for its calculation.

red codes post-adoption compared to the shelter was low (mean: 2.83; median: 0; sd: 7.51; range: 0 to 53). Most notably, at the first survey, approximately 10% of dogs (n = 24) had lower probabilities of green codes, and 10% of dogs (n = 24) and 13% (n = 31) higher chances of amber and red codes, for the *In kennel/House* context at the shelter, although this discrepency had largely disappeared by the second survey phone call. Around 20% of dogs (n = 53) had higher probabilities of green codes for the *Eating food* context at the shelter compared to the first survey, but only approximately 11% (n = 26) at the second survey. A smaller number of dogs (9 and 5 at first and second surveys, respectively) showed increased chances of red codes post-adoption for *Eating food* context.

3.6 Dog-level predictors

Length of stay at the shelter had a positive (higher latent scale scores), but weak relationship with dogs' average behaviour at the shelter ($\beta = 0.02$; 95% HDI: [0.01, 0.03]) and post-adoption ($\beta = 0.02$; 95% HDI: [0.00, 0.03]), and dogs neutered before arrival had higher latent scale scores on average than intact dogs ($\beta_{[\text{no}-\text{yes}]} = -0.38$; 95% HDI: [-0.73, -0.07]). Post-adoption, weight had a negative relationship (green codes increased with weight) with average behaviour ($\beta = -0.15$, 95% HDI: [-0.26, -0.05]).

4 Discussion

This study has provided the first comprehensive analysis of predicting shelter dog behaviour post-adoption from a longitudinal behavioural assessment. We applied a novel Bayesian hierarchical model to account for both shelter and post-adoption behaviour, missing data, and measurement error within a single statistical framework, which allowed making individual-level inferences of how dogs behaved in specific contexts at the shelter and in their new homes. We found few differences between how likely dogs were to show green, amber and red-coded behaviour on average at the shelter, and their corresponding chances of showing the same behavioural codes in the post-adoption surveys. Moreover, in contrast to some previous studies reporting high numbers of behavioural problems in shelter dogs (e.g. Gates et al. 2018), the behavioural reports in this study were largely dominated by green codes (i.e. the most favourable behaviours). Note, this does not imply that dogs behaved exactly the same post-adoption as they did in the shelter. Due to the varying amounts of data available per dog per context, there was considerable uncertainty in individual-level probabilities of green, amber and red codes, particularly in the post-adoption reports where dogs had a maximum of two records per context. Despite this, our overall results suggest that longitudinal behavioural data collected in shelters can offer reasonably accurate predictions of how dogs will behave post-adoption.

The hierarchical structure of our data allowed comparing variation in the behavioural records across dogs, across contexts, and across specific dogs $\times$ contexts combinations. Both at the shelter and post-adoption, the dog $\times$ context interaction explained the largest portion of behavioural variance (Figure 3a), and the correlations between shelter and post-adoption behaviour were only credibly larger than zero for the dog $\times$ context-varying intercept parameters (Figure 3b). Neither how dogs behaved on average across contexts, nor behaviour within contexts irrespective of particular dogs, were particularly illuminating for predicting behaviour post-adoption. Repeatability was also around 20% higher post-adoption than in the shelter for the dog $\times$ context behaviour, driven by greater heterogeneity in the post-adoption behaviour estimates than the shelter reports.

The positive correlation for dog behaviour within contexts at the shelter and post-

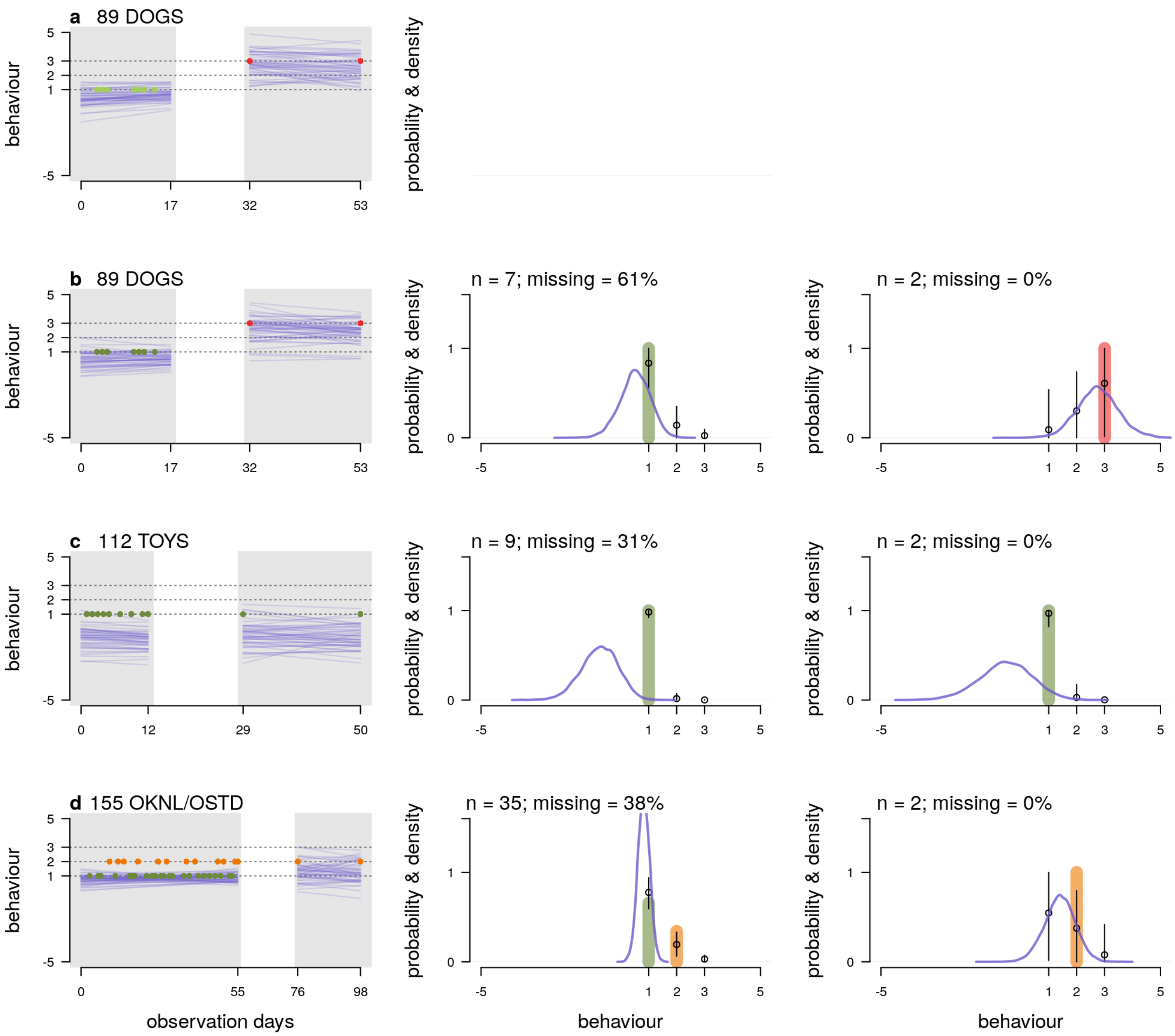

Figure 5: Posterior predictions for 4 specific dogs and contexts (a-d; anonymous dog ID number and context abbreviation shown above first panel of each plot). Left panels show behaviour over days after arrival to the shelter (first gray area) and behaviour over days after adoption (second gray area), where the coloured dots denote the raw ordinal data and blue lines 50 samples from the posterior distribution on the latent behavioural scale. Middle and right panels show the raw probability of green, amber and red codes (coloured bars) for the shelter (middle) and post-adoption (right), with the number of observations and percent missing data displayed above. Black dots and vertical line segments denote the mean and 95% highest density interval of predicted probabilities. Blue density curves show the underlying latent scale generating the ordinal codes for each dog.

adoption informs us about their linear association, but the practical implications of this positive correlation is not easy to discern. One option is to compute rates of false positives and negatives (e.g. Marder et al. 2013) or positive and negative predictive values (e.g. Patronek and Bradley 2016) to understand an assesment's predictive 'ability' (Patronek et al. 2019). This medical diagnostic setting lends itself easily to scenarios which have single shelter and post-adoption tests, and for tests where the outcome measure is a binary pass or fail result. In the results and Supplementary Materials, we provide certain crude diagnostics for the raw data. However, moving away from battery-style behavioural assessments in shelters necessitates moving away from labelling dogs as either problematic or not problematic, aggressive or non-aggressive, from limited behavioural information. Such sharp behavioural categorisations are at odds with the rationale for longitudinal assessments (e.g. ASPCA 2018; Rayment et al. 2015) that encourage forming a holistic profile of a dog's behaviour over different contexts, rather than focusing on any one single behaviour. Instead, we leveraged the posterior predictions of our model to determine how similar dog behaviour was across the post-adoption behaviour reports to each dog's behaviour on average at the shelter. Hierarchical models are excellently-suited to this task because they balance the information provided by each dog's data, on the one hand, and what is typical from the whole sample, on the other. Thus, any one dog's individual-level predictions are influenced by how many times certain behaviours occurred, how much missing data that dog has, and how typical their behaviour is for a specific context. This 'partial pooling' of information across dogs and contexts makes better predictions than non-hierarchical models or raw data because it balances statistically over- or under-fitting to the data (e.g. Gelman and Hill 2007; McElreath and Koster 2014).

To pick out a few representative examples (see the Supplementary Materials for more), Figure 5 illustrates the posterior predictions for four dogs, each in a different context, with the raw data for comparison. On the latent scale generating the ordinal data, dogs' predictions for post-adoption behaviour (indicated by the sample of regression lines in the left-hand plots) tend to be more uncertain than their in-shelter predictions, where more data is available overall. The impact of the amount of data available can also be seen by inspecting the overall probabilities of different code colours in the middle and right panels of each plot, where dogs with fewer data points have more uncertainty (e.g. the 95% HDIs in the middle panel of 5b). What occurs on average in certain contexts also affects the individual-level predictions. For instance, the dog in Figure 5b showed no red codes at the shelter for the *Interactions with dogs* context, but received red codes post-adoption. The overall increase in this dog's probability of red codes in the latter context compared to the shelter is 56%, but with a 95% highest density interval that spans from 100% to -1%, meaning there is a small chance that red codes were just as, or more, likely at the shelter than post-adoption. While this might seem surprising, it occurs because red codes were most probable at the shelter in the *Interactions with dogs* context overall, so even dogs that show no red codes in this context have an increased chance of red codes compared to other contexts. The *Interactions with dogs* context also had one of the highest amounts of missing data (dog 89 in Figure

5b has only 7 observations and 61% of missing data), and thus the predictions are more uncertain overall. Effectively, the model is telling us that it cannot be entirely sure that dog 89 has a greater tendency to show red codes post-adoption from the current data. By contrast, green codes were particularly dominant in the *Interactions with toys* context, and the amount of missing data was lower than average (Figure S1b), so dogs showing green codes in this context at the shelter and post-adoption have much tighter posterior estimates (e.g. see the predicted probabilities for the dog in Figure 5c).

For most contexts, the number of dogs whose 95% most likely values of the differences in probabilities of green, amber and red codes at the shelter compared to post-adoption did not include zero was low (Figure 4). This supports the ability for the longitudinal assessment to predict post-adoption behaviour, although we emphasise that this good predictive ability is in part due to the appropriate representation of uncertainty in the estimates, and the somewhat arbitrary, yet conventional, choice of 95% highest density interals (McElreath 2020). In the Supplementary Materials, we show that more dogs had 'significant' differences in their probabilities if we use the 89% highest density intervals. Regardless, the greatest discrepencies between shelter and post-adoption behaviour appeared in the *In kennel/House* and *Eating food* contexts. Between 10 and 15% of dogs in the former context were predicted to show more green codes, and fewer amber and red codes, at the first post-adoption phone call than at the shelter (Figure 4). This is perhaps unsurprising as the *In kennel* and *In House* contexts were the least comparable contexts across shelter and post-adoption periods, and a kennel would be expected to be a less stressful environment than the adoptive home. In contrast, around 20% of dogs at the first post-adoption survey, and around 11% at the second survey, showed lower chances of green codes than at the shelter in how they behaved around their food. The latter result is partly consistent with previous research questioning whether reasonable predictions of how dogs behave around food in the adopted home can be made from their behaviour at the shelter (e.g. Marder et al. 2013; Mohan-Gibbons et al. 2012). We also found that, post-adoption, the probability of red codes was highest in the *Eating food* context. However, the latter probability was still very low ($\sim 1\%$), and apart from nine dogs at the first survey and five dogs at the second, we found no evidence for increased probabilities of red codes post-adoption for behaviour around the food bowl. Thus, observational data collected on how dogs behave around their food bowl may provide a more valid indicator of post-adoption behaviour, even in this historically difficult-to-infer context.

Longitudinal behavioural assessments and adopter-reported behaviour are at risk of both missing data and measurement error, which are unlikely to be removed entirely. Instead, understanding when and how they emerge will be key to accounting for any deleterious effects they have on behavioural data. We found that dogs' odds of missing data at the shelter positively correlated with the amount of behavioural change across days after arrival, which means that dogs who did not change their behaviour over time were more likely to receive missing data on average at the shelter. This could have risen because dogs who change their behaviour less over time (i.e. less plastic) are not as high a priority for staff filling in behavioural reports in the time-constrained shelter environment. Importantly, there was no evidence to suggest that dogs with less favourable behaviour at the shelter were more likely to receive missing data (Table S3). In contrast, dogs with both higher chances of amber and red codes than average at the shelter, and dogs who changed less over time at the shelter, had lower chances of missing data post-adoption. Suprisingly, the correlation between a dog's average behaviour and missing data post-adoption was largely positive but still included zero (Table S3), meaning that adopter response rates were more strongly influenced by the dog's in-shelter behaviour. The exact mechanisms driving this correlation are unknown, so we urge it to be treated with caution at this stage.

Finally, the post-adoption behaviour reports had larger measurement error estimates than the shelter reports. Like estimates of zero-inflation in count data (Lambert 1992), we considered a 'green code inflation', that is, the proportion of green codes reported due to processes different to the main process generating the behavioural data (i.e. the dog's behaviour). In reality, the inflation component is a place-holder for more complex mechanisms, which include shelter staff not observing a dog's behaviour but inputting a green code anyway, or adopters reporting behaviours consistent with green codes, even when the real behaviour was less favourable. Green codes had approximately a 20% chance of being drawn from the inflation component of the model post-adoption, although there was large uncertainty (95% HDI: [3, 36]). By comparison, green code inflation was only 4% at the shelter. While we cannot disentangle the adopters' own behavioural reports from how the shelter behaviourists interpreted the behavioural descriptions from the adopters, this implies that systematic error in the adopter-reported behavioural reports is important to quantify further. Unfortunately, no previous studies have assessed patterns of measurement error in shelter or post-adoption behaviour.

5 Conclusion

Predicting the future behaviour of shelter dogs is a difficult task. Longitudinal assessments have been proposed as an alternative to battery-style, standardised tests, but research is still scarce on their implementation and efficacy. We have provided the first comprehensive analysis of a longitudinal assessment to predict behaviour post-adoption, and have presented a novel application of Bayesian hierarchical modelling to make statistical predictions of post-adoption behaviour. While dog behaviour was not fixed, few dogs showed marked changes in their behaviour post-adoption compared to the shelter, supporting the use of longitudinal assessments in shelters to inform adoption decisions, and predict post-adoption behaviour. We further demonstrated links between dog behaviour, missing data and measurement error, which should be accounted for in future analyses of longitudinal assessments.

References

ASPCA (2018). *Position statement on shelter dog behavior assessments*. `https://www.aspca.org/about-us/aspca-policy-and-position-statements/position-statement-shelter-dog-behavior-assessments`. Online; accessed 13-May-2020.

Bagozzi, B. E. and B. Mukherjee (2012). "A mixture model for middle category inflation in ordered survey responses". In: *Political Analysis* 20.3, pp. 369–386.

Bollen, K. S. and J. Horowitz (2008). "Behavioral evaluation and demographic information in the assessment of aggressiveness in shelter dogs". In: *Applied Animal Behaviour Science* 112.1-2, pp. 120–135.

Borsboom, D., G. J. Mellenbergh, and J. Van Heerden (2004). "The concept of validity." In: *Psychological review* 111.4, p. 1061.

Borsboom, D. et al. (2009). "The end of construct validity." In: *The Concept of Validity: Revisions, New Directions and Applications, Oct, 2008*. IAP Information Age Publishing.

Christensen, E. et al. (2007). "Aggressive behavior in adopted dogs that passed a temperament test". In: *Applied Animal Behaviour Science* 106.1-3, pp. 85–95.

Clay, L. et al. (2020). "In Defense of Canine Behavioral Assessments in Shelters: Outlining Their Positive Applications". In: *Journal of Veterinary Behavior*.

Cleasby, I. R., S. Nakagawa, and H. Schielzeth (2015). "Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance". In: *Methods in Ecology and Evolution* 6.1, pp. 27–37.

Dingemanse, N. J. et al. (2010). "Behavioural reaction norms: animal personality meets individual plasticity". In: *Trends in ecology & evolution* 25.2, pp. 81–89.

Gates, M. C. et al. (2018). "Post-adoption problem behaviours in adolescent and adult dogs rehomed through a New Zealand animal shelter". In: *Animals* 8.6, p. 93.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevelhierarchical models*. Vol. 1. Cambridge University Press New York, NY, USA.

Goold, C. and R. C. Newberry (2017a). "Modelling personality, plasticity and predictability in shelter dogs". In: *Royal Society open science* 4.9, p. 170618.

Goold, C. M. and R. C. Newberry (2017b). "Aggressiveness as a latent personality trait of domestic dogs: testing measurement invariance and local independence". In:

IntHout, J. et al. (2016). "Plea for routinely presenting prediction intervals in meta-analysis". In: *BMJ open* 6.7, e010247.

Kelley, M. E. and S. J. Anderson (2008). "Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model". In: *Statistics in medicine* 27.18, pp. 3674–3688.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kubinec, R. (2019). “Generalized Ideal Point Models for Time-Varying and Missing-Data Inference”. In:

Lambert, D. (1992). “Zero-inflated Poisson regression, with an application to defects in manufacturing”. In: *Technometrics* 34.1, pp. 1–14.

Marder, A. R. et al. (2013). “Food-related aggression in shelter dogs: A comparison of behavior identified by a behavior evaluation in the shelter and owner reports after adoption”. In: *Applied animal behaviour science* 148.1-2, pp. 150–156.

Marston, L. C. and P. C. Bennett (2003). “Reforging the bond—towards successful canine adoption”. In: *Applied Animal Behaviour Science* 83.3, pp. 227–245.

Maul, A., D. T. Irribarra, and M. Wilson (2016). “On the philosophical foundations of psychological measurement”. In: *Measurement* 79, pp. 311–320.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McElreath, R. and J. Koster (2014). “Using multilevel models to estimate variation in foraging returns”. In: *Human nature* 25.1, pp. 100–120.

Mohan-Gibbons, H., E. Weiss, and M. Slater (2012). “Preliminary investigation of food guarding behavior in shelter dogs in the United States”. In: *Animals* 2.3, pp. 331–346.

Mohan-Gibbons, H. et al. (2018). “The impact of excluding food guarding from a standardized behavioral canine assessment in animal shelters”. In: *Animals* 8.2, p. 27.

Mornement, K. et al. (2009). “Reliability, validity and feasibility of existing tests of canine behaviour”. In: *AIAM Annual Conference on Urban Animal Management, Proceedings*, pp. 11–18.

Mornement, K. M. et al. (2010). “A review of behavioral assessment protocols used by Australian animal shelters to determine the adoption suitability of dogs”. In: *Journal of Applied Animal Welfare Science* 13.4, pp. 314–329.

— (2014). “Development of the behavioural assessment for re-homing K9’s (BARK) protocol”. In: *Applied Animal Behaviour Science* 151, pp. 75–83.

Mornement, K. M. et al. (2015). “Evaluation of the predictive validity of the Behavioural Assessment for Re-homing K9’s (BARK) protocol and owner satisfaction with adopted dogs”. In: *Applied Animal Behaviour Science* 167, pp. 35–42.

Patronek, G. J. and J. Bradley (2016). “No better than flipping a coin: Reconsidering canine behavior evaluations in animal shelters”. In: *Journal of Veterinary Behavior* 15, pp. 66–77.

Patronek, G. J., J. Bradley, and E. Arps (2019). “What is the evidence for reliability and validity of behavior evaluations for shelter dogs? A prequel to “No better than flipping a coin””. In: *Journal of Veterinary Behavior*.

Poulsen, A., A. Lisle, and C. Phillips (2010). “An evaluation of a behaviour assessment to determine the suitability of shelter dogs for rehoming”. In: *Veterinary medicine international* 2010.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rayment, D. J. et al. (2015). "Applied personality assessment in domestic dogs: Limitations and caveats". In: *Applied Animal Behaviour Science* 163, pp. 1–18.

Stan Development Team et al. (2018). *CmdStan: the command-line interfact to Stan. Version 2.23.0.* URL: `http://mc-stan.org`.

Stan Development Team (2018). "Stan modeling language user's guide and reference manual, version 2.18. 0". In:

Taylor, K. D. and D. S. Mills (2006). "The development and assessment of temperament tests for adult companion dogs". In: *Journal of Veterinary Behavior* 1.3, pp. 94–108.

Tseng, C.-h. et al. (2016). "Longitudinal data analysis with non-ignorable missing data". In: *Statistical methods in medical research* 25.1, pp. 205–220.

Valsecchi, P. et al. (2011). "Temperament test for re-homed dogs validated through direct behavioral observation in shelter and home environment". In: *Journal of Veterinary Behavior* 6.3, pp. 161–177.

Van der Borg, J. A., W. J. Netto, and D. J. Planta (1991). "Behavioural testing of dogs in animal shelters to predict problem behaviour". In: *Applied Animal Behaviour Science* 32.2-3, pp. 237–251.

Vonesh, E. F., T. Greene, and M. D. Schluchter (2006). "Shared parameter models for the joint analysis of longitudinal data and event times". In: *Statistics in medicine* 25.1, pp. 143–163.

Wang, C.-C. and W.-C. Lee (2019). "A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals". In: *Research synthesis methods* 10.2, pp. 255–266.