Aggressiveness as a latent personality trait of domestic dogs: testing local independence and measurement invariance

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

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

[*] Corresponding author: conor.goold@nmbu.no (CG)

Abstract

Studies of animal personality attempt to uncover underlying or 'latent' personality traits that explain broad patterns of behaviour, often by applying latent variable statistical models (e.g. factor analysis) to multivariate data sets. Two integral, but infrequently confirmed, assumptions of popular latent variable models in animal personality are: i) behavioural variables are independent (i.e. uncorrelated) conditional on the latent personality traits they reflect (*local independence*), and ii) personality traits are associated with behavioural variables in the same way across individuals or groups of individuals (*measurement invariance*). We tested these assumptions using observations of aggression in four age classes (4 - 10 months, 10 months - 3 years, 3 - 6 years, over 6 years) of male and female shelter dogs (N = 4,743) in 11 different contexts. A structural equation model supported the hypothesis of two correlated (ρ = 0.25; p < 0.001) personality traits underlying aggression across contexts: aggressiveness towards people and aggressiveness towards dogs (comparative fit index: 0.97; Tucker-Lewis index: 0.96; root mean square error of approximation: 0.03). Aggression across contexts was moderately repeatable (towards people: ICC = 0.479, 95% CI: 0.466, 0.491; towards dogs: ICC = 0.303, 95% CI: 0.291, 0.315). However, certain contexts related to aggressiveness towards people (but not dogs) shared significant residual correlations unaccounted for by latent levels of aggressiveness. Furthermore, aggressiveness towards people and dogs in different contexts interacted with sex and age. Thus, sex and age differences in displays of aggression were not simple functions of underlying aggressiveness. Our results illustrate that the robustness of traits in latent variable

models must be critically assessed before making conclusions about the effects of, or factors influencing, animal personality. Our findings are of concern because inaccurate 'aggressive personality' trait attributions can be costly to dogs, recipients of aggression and society in general.

Key words: animal personality assessment; agonistic behaviour; shelter dogs; measurement bias; behavioural phenotyping

Introduction

Studies of non-human animal personality, defined by relatively consistent between-individual differences in behaviour, demonstrate that organisation of the behavioural phenotype is markedly hierarchical and non-independent [1–3]. Observed patterns of behaviour can be explained by broad behavioural dimensions or personality traits, including boldness, activity, exploration, sociability and aggressiveness [4], which are further inter-correlated to form behavioural syndromes [5]. To interpret the complexity inherent in behavioural phenotypes, personality traits and behavioural syndromes are frequently inferred using latent variable statistical models, which reduce two or more measured variables (the *manifest* variables) into one or more lower-order variables (the *latent* variables). Latent variable models have been popular in human psychology for over a century [6], and now comprise a flexible set of methods to derive lower-order variables from multivariate data sets [7].

Because latent variables are unobserved, latent variable models require careful application and interpretation [8–10]. In animal personality, many studies use *formative* models, such as principal components analysis, that construct composite variables comprised of linear combinations of measured variables. However, formative models impose only weak assumptions about the relationships between latent variables and measured variables [11]. As such, their utility for inferring the nature of personality traits has been criticised in both animals [12,13] and humans [9,11,14,15]. Alternatively,

researchers have used *reflective* models, such as factor analysis and, increasingly, confirmatory approaches such as structural equation modelling. These models regress measured behaviours on one or more latent variables [11], incorporating measurement error and possibilities to compare *a priori* competing hypotheses (for applications in animal personality, see [1,16–18]).

Whilst reflective models offer a powerful framework to probe the latent variable structure of behaviour, they impose certain constraints on the interpretation and modelling of latent variables that have received scrutiny in human psychology but are rarely discussed in studies of animal personality. Two foundational assumptions of these models are *local independence* and *measurement invariance*. Local independence implies that manifest variables should be independent of each other conditional on the latent variables [19,20]. For example, given a latent variable θ and two binary manifest variables Y_1 and Y_2 that can take the values 0 and 1, the item response theory model asserts that $P(Y_1 = 1, Y_2 = 1|\theta) = P(Y_1 = 1|\theta)P(Y_2 = 1|\theta)$. As such, the latent variables should screen off any dependence between manifest variables. Measurement invariance, on the other hand, implies that the latent variables function the same (i.e. are invariant or equivalent) in different subsets of a population or in the same individuals through time [21–24]. That is, given a fixed level x on a latent variable θ, denoted θ_x, the expected values of manifest variables Y_i, from *i = 1* to *N*, should be the same across a grouping factor π (e.g. sex or different populations), expressed as $E(Y_i \mid \theta_x) = E(Y_i \mid \theta_x, \pi)$. Intuitively, violations of either local independence or

measurement invariance imply that the latent variables do not completely explain variation in manifest variables, which may lead to artefactual conclusions about the differences between individuals as a function of trait scores [25–27]. More generally, investigating violations of local independence and measurement invariance can identify measurement biases that can be rectified to improve the quality of personality assessments.

The use of latent variable modelling is particularly common in studies of personality in domestic dogs (*Canis lupus familiaris*), where the collection of data sets comprised of multiple behavioural variables (e.g. questionnaires) has been prevalent for decades [28–30]. Personality assessments are important both in applied settings to predict the behaviour of dogs at future time points [31] and also to elucidate behavioural traits pertinent to dogs' domestication history [32,33]. Decomposing behaviour into a smaller number of underlying dimensions that explain variation in measured behaviours aids in measuring hypothesised stable behavioural features of individuals from which predictions of future behaviour are estimable and upon which selection may have been focused. Nonetheless, research on personality in dogs has led to different numbers and composition of hypothesised personality traits with little clear consensus on how such traits should be compared within and between studies [34–36]. Most importantly, the predictive value of personality assessments in dogs has been inconsistent [31,37–41], perhaps most prominently in shelter dog personality assessments (e.g. see [31] for a review). The predictive validity of personality assessments is of particular concern in

102 tests of aggression, where aggression is frequently divided into different aggressiveness
103 traits, including owner-, stranger-, dog- or animal-directed factors [29,35,42,43]. Based
104 on such assessments, if falsely labelled as aggressive in one or more of these categories,
105 the dog may be euthanised unnecessarily whereas, if falsely considered unaggressive,
106 the dog may be more likely to be placed in situations leading to harmful behaviour. Dog
107 bites are a serious public health concern [44], especially for animal shelters rehoming
108 dogs to new owners, and aggressive behaviour is undesirable to many organisations
109 using dogs for various working roles [45].

110

111 Why has the prediction of dog personality been difficult? On one hand, the validation of
112 personality tests has been criticised for lacking rigour [34,46], meaning the targeted
113 traits may not be under measurement as intended. In particular, commonly-used
114 'battery assessments', comprised of different sub-tests conducted sequentially, may not
115 be valid if the responses to one sub-test alter how the animal responds to subsequent
116 sub-tests given an underlying trait (e.g. if test items are 'invasive' to subsequent items;
117 [47]). Moreover, test batteries can be "inherently stress-inducing for test animals" ([31]:
118 10), so behavioural responses may be functions of the targeted traits and stress
119 responses. On the other hand, personality may simply not be generalisable across all
120 individuals if individual behaviour is differentially dependent on factors such as age (e.g.
121 [35,40]) or fluctuating environmental conditions (e.g. behaviour in a shelter versus out
122 of a shelter; [48]). In reflective latent variable modelling, greater attention to the
123 assumptions of local independence and measurement invariance is, thus, warranted.

124 For aggressiveness specifically, if aggression shown in one context or test item increases
125 the likelihood of aggression in others [48], local independence would be violated.
126 Further, interactive effects between aggressiveness and demographic variables, such as
127 age, sex, breed or neuter status (e.g. [49–51]), would violate the assumption of
128 measurement invariance. Whilst van den Berg *et al.* [18] reported measurement
129 invariance of the *stranger-directed aggression* factor from the Canine Behaviour and
130 Research Questionnaire (C-BARQ; [29]) across three breeds of dogs, measurement
131 invariance of aggression assessed with respect to other factors has not been evaluated
132 and, to our knowledge, no authors have confirmed local independence between
133 manifest variables in animal personality assessments.

134

135 In this paper, we assessed local independence and measurement invariance of
136 aggressiveness using a large sample of data on inter-context aggression in shelter dogs.
137 First, we sought to decompose observations of aggression towards people and dogs
138 across contexts into separate aggressiveness traits. Secondly, we assessed whether
139 aggression in different contexts remained associated beyond that explained by latent
140 levels of aggressiveness, testing local independence. Thirdly, we investigated whether
141 the probability of aggression in different contexts believed to be underpinned by the
142 same aggressiveness trait was measurement invariant with respect to sex and age
143 groups. While it may be unrealistic for measurement invariance to hold in all instances,
144 it is important to establish whether it holds across basic biological variables such as age

and sex which are generally applicable to dog populations undergoing personality assessment.

Materials & Methods

Subjects

Observational data on the occurrence of aggression in 4,743 dogs were gathered from Battersea Dogs and Cats Home's (UK) observational and longitudinal dog behaviour assessment records. The data were from a sample of dogs (N=4,990) at the shelter's three rehoming centres during 2014 (including dogs that arrived during 2013 or left in 2015). We selected the records from all dogs that were at least 4 months old, excluding younger dogs because they were more likely to be unvaccinated, more limited in their interactions at the shelter and may have been kennelled in different areas to older dogs. Relevant demographic characteristics of the dogs are provided in Table 1. Although dogs were from a variety of heritages (including purebreeds and mongrels), the analyses here did not explore breed differences because visual assessment to attribute breed to dogs with unknown heritage has been questioned [52].

162

Table 1. Demographic characteristics of the studied dogs.

Variable	**Mean ± SD / *N***
Average age at shelter (years; all ≥ 4 months of age)	3.75 ± 3.03
Total days at the shelter	25.13 ± 41.53
Weight (average weight if multiple measurements; kg)	19.06 ± 10.26
Rehoming centre: London / Old Windsor / Brands Hatch	2897 / 1280 / 566
Males / females	2749 / 1994
Neutered before arrival / neutered at shelter / not neutered	1218 / 1665 / 1502
Relinquished by owners / returned to shelter / strays	2892 / 260 / 1591

163

164 ## Shelter environment

165 The shelter was composed of three different UK rehoming centres: Battersea, London
166 (average capacity in 2014/2015: 167 dogs); Old Windsor (average capacity in
167 2014/2015: 133 dogs); Brands Hatch (average capacity in 2014/2015: 47 dogs). All dogs
168 arrived in an intake area of their respective rehoming centre and, when considered
169 suitable for adoption, were moved to a 'rehoming' area that was partially open to the
170 public between 1000 h and 1600 h. All kennels were indoors. Kennels varied in size, but
171 were usually approximately 4m x 2m and included either a shelf and bedding alcove
172 area, or a more secluded bedding area at the back of the kennel (see [53] for more
173 details). At different times throughout the day, dogs had access to indoor runs behind
174 their kennels. In each kennel block area, dogs were cared for (e.g. fed, exercised, kennel
175 cleaned) by a relatively stable group of staff members, allowing the development of

176 familiarity with staff members and offering some predictability for dogs after arrival at
177 the shelter. Although data on the number of dogs in each kennel were incomplete, in
178 the majority of cases dogs were kennelled singly for safety reasons. The shelter mainly
179 operated between 0800 h and 1700 h each day. All dogs were socialised with staff
180 and/or volunteers each day (often multiple times) unless it was unsafe to do so. They
181 were provided water ad libitum and fed commercial complete dry and/or wet tinned
182 food twice daily (depending on recommendations by veterinary staff). Dogs received
183 daily tactile, olfactory and/or auditory enrichment/variety (e.g. toys, essential oils,
184 classical music, time in a quiet 'chill-out' room).

185

186 Data collection

187 In the observational assessment procedure, trained shelter employees recorded
188 observations of dog behaviour in a variety of contexts as part of normal shelter
189 procedures. The average number of days between successive observations within all
190 contexts and across all dogs was 3.27 (SD = 2.08), and dogs had an average of 9.77 (SD =
191 13.41) observations within each context. Behavioural observations pertaining to each
192 context were completed using an ethogram specific to that context and recorded in a
193 custom computer system. Multiple observations could be completed each day. The
194 ethogram behaviour that best described a dog's behaviour in a particular context during
195 an observation was recorded by selecting it from a series of drop-down boxes (one for
196 each context). Although staff could also add additional information in character fields,

we did not analyse those comments in this study. The ethogram for each context represented a scale of behaviours ranging from desirable to undesirable considered by the shelter to be relevant to dog welfare and ease of adoption. Contexts had between 10 and 16 possible behaviours to choose from, some of which overlapped between different contexts. Among the least desirable behaviours in each context was aggression towards either people or dogs (depending on context). Aggression was formally defined as "*Growls, snarls, shows teeth and/or snaps when seeing/meeting other people/dogs, potentially pulling or lunging towards them*", distinguished from non-aggressive but reactive responses, defined as "*Barks, whines, howls and/or play growls when seeing/meeting other people/dogs, potentially pulling or lunging towards them*". Observation contexts included both onsite (at the shelter) and offsite (e.g. out in public parks) settings.

For the analyses here, we focused on observations of aggression in nine core onsite contexts: i) *Handling*, ii) *In kennel*, iii) *Out of kennel*, iv) *Interactions with familiar people*, v) *Interactions with unfamiliar people*, vi) *Eating food*, vii) *Interactions with toys*, viii) *Interactions with female dogs*, ix) *Interactions with male dogs*. For the *In kennel* and *Out of kennel* contexts, recording of aggression towards both people and dogs was possible. Although multiple observations could be made, if both occurred at the same time, aggression towards people would be prioritised over aggression towards dogs in the recording process. Therefore, *In kennel* and *Out of kennel* were each divided to reflect aggression shown towards people and dogs, respectively. This resulted in 11 final

aggression contexts, defined in Table 2, which were used as manifest variables in structural equation models to investigate latent aggressiveness traits. Each observation of aggression was recorded in the category that best described the scenario. Nonetheless, certain contexts could occur closely in space and time. For example, the *Handling* context could directly succeed the *Interactions with familiar people* or *Interactions with unfamiliar people* contexts. The sequential occurrence of certain contexts was used to inform tests of local independence, explained below.

We aggregated behavioural observations across time for each dog into a dichotomous variable indicating whether a dog had or had not shown aggression in a particular context at any time while at the shelter. This was performed because the overall prevalence of aggression was low, with only 1.06% of all observations involving aggression towards people and 1.13% towards dogs. Thus, the main difference between individuals was whether they had or had not shown aggression in a particular context during their time at the shelter (see Table S1 for raw counts of aggression by context), and we interpret aggressiveness here as a between-individual difference variable.

Table 2. Behavioural observation contexts.

Context	Definition
Handling	A dog's reaction to informal handling by people (e.g. stroking non-sensitive areas, touching the collar, fitting a harness or lead).
In kennel towards people	A dog's reaction to people approaching or walking past the kennel.
In kennel towards dogs	A dog's reaction to dogs in neighbouring kennels or dogs walking past the kennel.
Interactions with familiar people	A dog's reaction when outside of the kennel to familiar people (interacted with at least once before) approaching, making eye contact, speaking to or attempting to make physical contact with the dog.
Interactions with unfamiliar people	A dog's reaction when outside of the kennel to unfamiliar people (never interacted with before) approaching, making eye contact, speaking to or attempting to make physical contact with the dog.
Out of kennel towards people	A dog's reaction when around people outside of the kennel. Large distances may separate the focal dog and people, and no attempt is made to engage by the people with the focal dog.
Out of kennel towards dogs	A dog's reaction when around dogs outside of the kennel. Large distances may separate the focal dog and other dogs, and the dogs are not encouraged to interact.
Eating food	A dog's reaction when eating food (e.g. from a food bowl, or toy filled with food) to people approaching, in close proximity, or attempting to touch the food container.
Interactions with toys	A dog's reaction when interacting with toys to people approaching within close proximity and/or attempting to touch the toy.
Interactions with female dogs	A dog's reaction during structured interactions with a female dog, including approaching each other, walking in parallel, and interacting off-lead. Both dogs are aware of each other's presence and are in close enough proximity to engage in a physical interaction.
Interactions with male dogs	A dog's reaction during structured interactions with a male dog, including approaching each other, walking in parallel, and interacting off-lead. Both dogs are aware of each other's presence and are in close enough proximity to engage in a physical interaction.

Behavioural observation contexts analysed for the presence or absence of aggression.

Validity of behaviour recordings

Validity of the recording of behaviour was assessed separately from the main data collection as part of a wider project investigating the use of the observational assessment method. Ninety-three shelter employees trained in conducting behavioural observations each watched (in groups of 5 – 10 people) 14 videos, approximately 30 seconds each, presenting exemplars of 2 different behaviours from seven contexts (to keep the sessions concise and maximise the number of participants). For each context, behaviours were chosen pseudo-randomly by numbering each behaviour and selecting two numbers using a random number generator. Experienced behaviourists working at the shelter filmed the videos demonstrating the behaviours. Videos were shown to participants once in a pseudo-random order. After each video, participants recorded on a paper answer sheet the behaviour they thought most accurately described the dog's behaviour based on the ethogram specific to the context depicted. Two of the videos illustrated aggression: one in a combined *Interactions with new* and *familiar people* context (combined because familiarity between specific people and dogs was not universally known) and one in the *In kennel towards dogs* context. The authors were blind to the selection of videos shown and to the video coding sessions with shelter employees.

Data analysis

255 All data analysis was conducted in R version 3.3.2 [54].

256

257 Validity of behaviour recordings

258 The validity of shelter employees' recording of behaviour from videos was assessed by
259 the percentage of participants who identified the 2 videos as showing examples of
260 aggression.

261

262 Missing data

263 Data were missing when dogs did not experience particular contexts while at the
264 shelter. The missing data rate was between 0.06% and 5% for each context, except for
265 the *Interactions with female dogs* and *Interactions with male dogs* categories which had
266 17% and 18% of missing values, respectively (because structured interactions with other
267 dogs did not arise as frequently). Moreover, 16% and 7% of dogs were missing weight
268 measurement and neuter status data, respectively, which were independent variables
269 statistically controlled for in subsequent analyses. We created 5 multiply imputed data
270 sets [55], upon which all of the analyses in the sections below were conducted and
271 results pooled. The multiple imputation took into account the hierarchical structure of
272 the data (observations within dogs), all independent variables reported below, and the
273 data types (ordered binary variables for the context data, positive-continuous for weight

measurements, nominal for neuter status). The data were assumed to be missing at random, that is, dependent only on other variables in the analyses.

Structural equation models

We used structural equation modelling to assess whether aggression towards people (contexts: *Handling*, *In kennel towards people*, *Out of kennel towards people*, *Interactions with familiar people*, *Interactions with unfamiliar people*, *Eating food*, *Interactions with toys*) and towards dogs (contexts: *In kennel towards dogs*, *Out of kennel towards dogs, Interactions with female dogs*, *Interactions with male dogs*) could be explained by two latent aggressiveness traits: aggressiveness towards people and dogs, respectively. We compared a model where the latent variables were orthogonal to a model where the latent variables were allowed to correlate, since positive correlations between different aggressiveness traits in dogs have been reported in dogs [56]. Models were fit using the *lavaan* package [57], with the weighted least squares mean and variance adjusted (WLSMV) estimator and theta/conditional parameterisation, as recommended for categorical dependent variables [7,58,59]. The latent variables were standardised to have mean 0 and variance 1. The results were combined across imputed data sets using the 'runMI' function in the *semTools* package [60]. The fit of each model was ascertained using the comparative fit index (CFI) and Tucker Lewis index (TLI), where values > 0.95 indicate excellent fit, as well as the root mean squared error of

approximation (RMSEA) where values < 0.06 indicate good fit [7]. Parameter estimates were summarised by *p* values (considered significant at $p < 0.05$).

Local independence

We tested the assumption of local independence by re-fitting the best-fitting structural equation model with residual correlations between context variables. To maintain a theoretically driven approach (see [61] regarding the best practice of including residual correlations in structural equation models) and model identifiability, we only tested a predefined set of residual correlations that we believed to be most relevant. First, we allowed correlations between *Handling* with *In kennel towards people*, *Interactions with familiar people*, *Interactions with unfamiliar people* and *Interactions with toys*, respectively. The *Handling* context could directly succeed these other contexts, leading to close temporal-spatial relationships, and whether a dog showed aggression in the *Handling* context may be mediated by a person's decision to handle a dog depending on the dog's behaviour in preceding contexts. The residual correlation between *Handling* and *Eating food* was not estimated because shelter employees would be unlikely to handle a dog while the dog ate its daily meals. The residual correlation between *Handling* and *Out of kennel towards people* was not estimated because any association between *Handling* and *Out of kennel towards people* would be mediated by either the *Interactions with familiar people* or *Interactions with unfamiliar people* context. Therefore, secondly, we estimated the three-way correlations between *Out of kennel*

towards people, *Interactions with familiar people* and *Interactions with unfamiliar people*. Similarly, and lastly, we estimated the three-way correlations between *Out of kennel towards dogs, Interactions with female dogs* and *Interactions with male dogs*.

Measurement invariance

To test for measurement invariance in each of the latent traits derived from the best fitting structural equation model, we investigated the response patterns across aggression contexts related to the same latent aggressiveness trait using Bayesian hierarchical logistic regression models. Whilst measurement invariance for categorical data can be ascertained in structural equation model frameworks [58,62], item response theory is more commonly applied to dichotomously scored variables. In psychometrics, the 1-parameter item response theory model, or Rasch model, represents the probability that an individual responds correctly to a particular test item as a logistic function of i) each individual's latent ability and ii) the item's difficulty level. The Rasch model can be expressed as a hierarchical logistic regression model [63,64], whereby individual latent abilities are modelled as individual-specific intercepts (i.e. 'random intercepts'), the propensity for a correct answer to an item i is its regression coefficient β_i, and credible interactions between items and relevant independent variables (e.g. group status) indicate a violation of measurement invariance. Here, the dependent variable was the binary score for whether or not dogs had shown aggression in each context and the average probability of aggression across contexts varied by dog,

representing latent levels of aggressiveness. Context type, dog age, dog sex and their interactions were included as categorical independent variables.

Age was treated as a categorical variable, with categories reflecting general developmental periods: i) 4 months to 10 months (juvenile dogs before puberty), ii) 10 months to 3 years (dogs maturing from juveniles to adults), iii) 3 years to 6 years (adults), and iv) 6 years + (older dogs). Broad age categories were chosen due to potentially large differences in developmental timing between individuals. Age was categorised because we predicted that aggression would be dependent on these developmental periods.

Models included additional demographic variables (see Table 1) that may mediate the probability of aggression: body weight (average weight if multiple measurements were taken), total number of days spent at the shelter, the rehoming centre at which dogs were based (London, Old Windsor, Brands Hatch), neuter status (neutered before arrival, neutered at the shelter, not neutered) and source type (relinquished by owner, returned to the shelter after adoption, stray). Categorical variables were represented as sum-to-zero deflections from the group-level intercept to ensure the intercept represented the average probability of aggression across categorical levels. Weight and total days at the shelter were mean-centered and standardised by 2 standard deviations. Due to the potentially complex relationships between these variables and

aggression (e.g. interactive effects between neuter status and sex; [49]), which could also include violations of measurement invariance, we decided not to interpret their effects inferentially. Instead, they were included to make the assessment of measurement invariance between sexes and age groups conditional on variance explained by potentially important factors.

For comparability to other studies in animal personality, behavioural repeatability was calculated across contexts in each model using the intraclass correlation coefficient (ICC), calculated as $\frac{\sigma_\beta^2}{\sigma_\beta^2+\sigma_\epsilon^2}$, where σ_β^2 represented the between-individual variance of the probability of aggression (i.e. the variance of the random intercepts), and σ_ϵ^2 was $\pi^2/3$, the residual variance of the standard logistic distribution [65].

Computation

Models were computed using the probabilistic programming language Stan version 2.12 [66], using Hamiltonian Monte Carlo, a type of Markov Chain Monte Carlo (MCMC) algorithm, to sample from the posterior distribution (model code supplied in Supporting Information). Prior distributions for all independent variables were normal distributions with mean 0 and standard deviation 1, attenuating regression coefficients towards zero for conservative inference. The prior on the overall intercept parameter was normally distributed with mean 0 and standard deviation 5. The standard deviation of dog-

377 specific intercept parameters was given a half-Cauchy prior distribution with mean 0
378 and shape 2. Each model was run with 4 chains of 2,000 iterations with a 1,000 step
379 warm-up period. The Gelman-Rubin statistic (ideally < 1.05) and visual assessment of
380 traceplots were used to assess MCMC convergence and we checked the accuracy of the
381 posterior predicted probabilities of aggression against the raw data. Regression
382 coefficients were expressed as odds ratios and were summarised by their mean and 95%
383 Bayesian credibility interval (CI, i.e. the 95% most probable parameter values). To
384 compare levels of categorical variables and their interactions, we computed the 95% CI
385 of the differences between the respective posterior distributions.

386

387 ***Model selection & parameter inference***

388 Models were run on each imputed data set and their respective posterior distributions
389 were averaged to attain a single posterior distribution for inference. Adopting a
390 Bayesian approach allowed the estimation of interaction parameters (i.e. testing
391 measurement invariance) without requiring corrections for multiple comparisons as in
392 frequentist null hypothesis testing [67]. Nonetheless, models included a large number of
393 estimated parameters. Two strategies were employed to guard against over-fitting of
394 models to data. First, we selected the model with the best out-of-sample predictive
395 accuracy given the number of parameters based on the Widely Applicable Information
396 Criterion (WAIC; using the R package *loo* [68]). Four variants of each model were
397 computed: two-way interactions between contexts and age and contexts and sex,

398 respectively (model 1), a single interaction with sex but not with age (model 2), a single
399 interaction with age but not with sex (model 3), and no interactions (model 4). All
400 models included the mediating independent variables above. Second, to avoid testing
401 point-estimate null hypotheses, the effect of a parameter was only considered credibly
402 different from zero if the odds ratio exceeded the region of practical equivalence (ROPE;
403 see [69]) around an odds ratio of 1 from 0.80 to 1.25. An odds ratio of 0.80 or 1.25
404 indicates a 20% decrease or increase (i.e. 4/5 or 5/4 odds), respectively, in the odds of
405 an outcome, frequently used in areas of bioequivalence testing (e.g. [70]), which we
406 here considered to be small enough to demonstrate a negligible effect in the absence of
407 additional information. If a 95% CI fell completely within the ROPE, the null hypothesis
408 of no credible influence of that parameter was accepted; if a 95% CI spanned the ROPE
409 (i.e. included part of the ROPE), then the parameter's influence was left undecided [69].

410

411 **Ethics statement**

412 Permission to access and utilise the data was given by the shelter after signing a non-
413 disclosure agreement. Approval from an ethical review board was not required for this
414 study.

Data accessibility

The data used in this study are protected by a non-disclosure agreement. Researchers can access the data by contacting Battersea Dogs and Cats Home.

Results

Validity of behaviour recordings

For the video showing aggression towards people, 51.61% of participants identified the behaviour as aggression and 41.94% identified the behaviour as non-aggressive but reactive behaviour (see definitions above). For the video showing aggression towards dogs, 52.69% identified the behaviour correctly and 44.09% identified the behaviour as non-aggressive but reactive behaviour. For the 12 other videos not showing aggression, only 1 person coded a video as aggression towards people and 3 people coded videos as aggression towards dogs.

Structural equation models

The raw tetrachoric correlation matrix for aggression between contexts is presented in Table S2. Both structural equation models demonstrated excellent fit, with the model

433 with correlated latent variables fitting marginally better (CFI: 0.97; TLI: 0.96; RMSEA:
434 0.03) than the model with uncorrelated variables (CFI: 0.96; TLI: 0.95; RMSEA 0.04). All
435 regression coefficients of the model with correlated latent variables were positive and
436 significant (Table 3), and latent variables shared a significant positive correlation (ρ =
437 0.25; $p < 0.001$; Table 3).

438

439 Local independence

440 Allowing the pre-defined residuals to co-vary in the structural equation model resulted
441 in marginally better fit (CFI = 0.98; TLI = 0.97; RMSEA: 0.03). A significant negative
442 correlation was observed between *Handling* and *In kennel towards people* contexts (ρ =
443 -0.67; p = 0.003; Table 4), and *Handling* and *Interactions with unfamiliar people* contexts
444 (ρ = -0.55; p = 0.01; Table 4). Significant positive correlations were observed between
445 *Handling* and *Interactions with toys* contexts (ρ = 0.15; p = 0.04; Table 4), and *Out of*
446 *kennel towards people* with the *Interactions with unfamiliar people* context (ρ = 0.27; p
447 < 0.001; Table 4). No significant residual correlations between contexts reflecting
448 aggressiveness towards dogs were observed (Table 4).

449

450

Table 3. Structural equation model parameter estimates.

Parameter	Estimate	SE	*z* value	*p* value
Handling[a]	0.83	0.06	15.07	< 0.001
In kennel towards people[a]	1.31	0.09	15.00	< 0.001
Out of kennel towards people[a]	0.33	0.07	12.60	< 0.001
Interactions with familiar people[a]	0.99	0.07	14.81	< 0.001
Interactions with unfamiliar people[a]	1.58	0.12	12.97	< 0.001
Eating food[a]	0.73	0.06	13.02	< 0.001
Interactions with toys[a]	0.54	0.07	8.04	< 0.001
In kennel towards dogs[b]	0.75	0.06	11.74	< 0.001
Out of kennel towards dogs[b]	0.50	0.04	11.87	< 0.001
Interactions with female dogs[b]	0.94	0.07	11.19	< 0.001
Interactions with male dogs[b]	0.89	0.07	12.66	< 0.001
Correlation: People ~ Dogs	0.25	0.03	7.50	< 0.001

Parameter estimates from the best-fitting structural equation model, explaining aggression by two correlated latent variables.

[a] Contexts reflecting aggressiveness towards people

[b] Contexts reflecting aggressiveness towards dogs

451

452

Table 4. Tests of local independence.

Residual correlations	Estimate	SE	*z* value	*p* value
Handling ~ In kennel towards people[a]	-0.67	0.22	-3.01	0.003
Handling ~ Interactions with familiar people[a]	0.15	0.09	1.72	0.09
Handling ~ Interactions with unfamiliar people[a]	-0.55	0.21	-2.66	0.01
Handling ~ Interactions with toys[a]	0.15	0.07	2.06	0.04
Out of kennel towards people ~ Interactions with familiar people[a]	0.04	0.08	0.51	0.61
Out of kennel towards people ~ Interactions with unfamiliar people[a]	0.27	0.09	3.20	0.001
Interactions with familiar people ~ Interactions with unfamiliar people[a]	0.003	0.11	0.02	0.98
Out of kennel towards dogs ~ Interactions with female dogs[b]	-0.65	0.63	-1.04	0.30
Out of kennel towards dogs ~ Interactions with male dogs[b]	-0.41	0.67	-1.08	0.27
Interactions with female dogs ~ Interactions with male dogs[b]	-0.28	0.57	-0.50	0.62

Estimated residual correlations between *a priori* defined structural equation model parameters.

[a] Contexts reflecting aggressiveness towards people

[b] Contexts reflecting aggressiveness towards dogs

453

454 ## Measurement invariance

455 Separate models were run for contexts reflecting aggressiveness towards people and

456 aggressiveness towards dogs. For all models, all Gelman-Rubin statistics were all < 1.01,

457 effective sample sizes for all parameters were > 1000, and traceplots showed good
458 mixing. Posterior predictive checks of model estimates reflected the raw data (Figs 1
459 and 2). The full measurement invariance model (model 1) including interactions
460 between contexts and sex and contexts and age groups had the best out-of-sample
461 predictive accuracy for both the aggressiveness towards people and aggressiveness
462 towards dog models, respectively, illustrated by the lowest WAIC values (Table 5). Since
463 some models included numerous interactions, we provide an overall summary of the
464 main results below (Figs 1 and 2) with full parameter estimates provided in Tables S3
465 and S4.

Table 5. Bayesian hierarchical model selection using WAIC.

Model	Aggressiveness towards people	Aggressiveness towards dogs
Model 1	13405.6 ± 179.0	15257.2 ± 133.1
Model 2	13506.3 ± 179.6	15381.4 ± 133.4
Model 3	13426.3 ± 179.1	15285.3 ± 133.0
Model 4	13521.7 ± 179.5	15407.6 ± 133.4

Mean ± standard error of the Widely Applicable Information Criteria (WAIC) values per model and latent variable (aggressiveness towards people and dogs, respectively), used to assess measurement invariance: model 1 (age x context and sex x context interactions), model 2 (sex x context interaction only), model 3 (age x context interaction only), model 4 (no interactions). Models with the lowest WAIC values are estimated to have the best out-of-sample predictive accuracy.

466

467 **Aggressiveness towards people**

468 The odds of aggression towards people, across categorical predictors and for an average
469 dog of mean weight and length of stay at the shelter, were 0.022 (CI: 0.021 to 0.024), a

470 probability of approximately 2%. On average, aggression was most likely in the *In kennel*
471 *towards people* context (OR = 0.054; CI: 0.049 to 0.058) and least probable in the
472 *Interactions with toys* context (OR = 0.008; CI: 0.007 to 0.009).

473

474 Aggression was less likely across contexts for females than males (OR = 0.719; CI: 0.668
475 to 0.770), although there were also credible interactions between sex and contexts (Fig
476 1a; Table S3). Whereas males and females had similar odds of aggression in the *Out of*
477 *kennel towards people* context, smaller differences were observed between *Out of*
478 *kennel towards people* and *Handling* (OR = 0.578; CI: 0.481 to 0.682), *Eating food* (OR =
479 1.812; CI: 1.495 to 2.152) and *Interactions with familiar people* (OR = 1.798; CI: 1.488 to
480 2.126) contexts in females compared to males. Aggression in the *Interactions with*
481 *unfamiliar people* context was also similar between males and females, while larger
482 differences were observed between *Interactions with unfamiliar people* and *Handling*
483 (OR = 0.616; CI: 0.530 to 0.702), *Eating food* (OR = 0.594; CI: 0.506 to 0.686) and
484 *Interactions with familiar people* (OR = 0.598; CI: 0.513 to 0.687) contexts in females
485 compared to males.

486

487 Apart from lower odds of aggression in 4 to 10 month olds compared to 10 month to 3
488 year old dogs (OR = 0.638; CI: 0.565 to 0.705), there was no simple influence of age
489 group on aggressiveness. Between the 4 to 10 months old and 3 to 6 years old groups,
490 differences between the odds of aggression across contexts varied due to an increase of

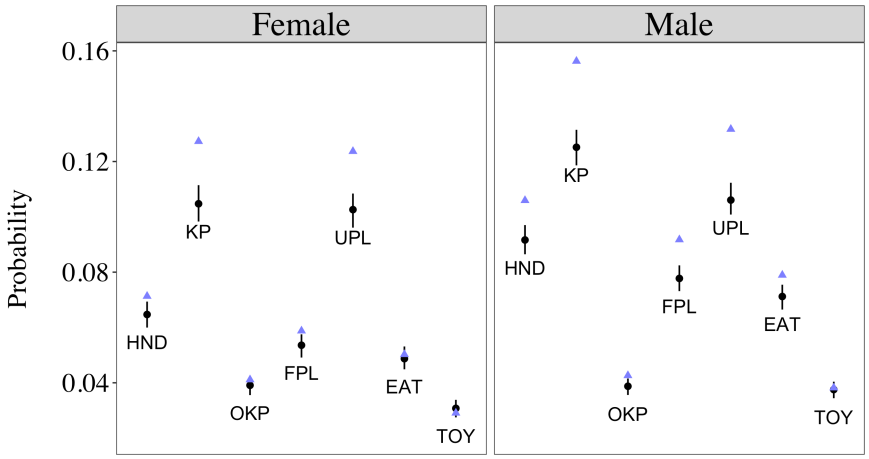
aggression in certain contexts but not others (Fig 1b; Table S4). Aggression in *In kennel towards people* and *Interactions with unfamiliar people* contexts particularly increased, leading to larger differences between, for example, *In kennel towards people* and *Eating food* (OR = 0.524; CI: 0.400 to 0.642) and *Eating food* and *Interactions with unfamiliar people* (OR = 1.721; CI: 1.403 to 2.059) contexts for 10 month to 3 year olds compared to 4 to 10 month olds, and between *In kennel towards people* and *Out of kennel towards people* (OR = 0.470; CI: 0.355 to 0.606) and *Out of kennel towards people* and *Interactions with unfamiliar people* (OR = 2.051; CI: 1.608 to 2.543) contexts in 3 to 6 year olds compared to 4 to 10 month olds. In 3 to 6 year old compared to 10 month to 3 year old dogs, aggression increased in the *Handling* and *Eating food* contexts but decreased in the *Out of kennel towards people* context, resulting in larger differences between, for instance, *Handling* and *Out of kennel towards people* (OR = 0.526; CI: 0.409 to 0.631) and *Out of kennel towards people* and *Interactions with unfamiliar people* (OR = 2.349; CI: 1.891 to 2.925), and smaller differences between *Eating food* and *Interactions with familiar people* (OR = 0.576; CI: 0.468 to 0.687).

Dogs over 6 years old demonstrated qualitatively different response patterns across certain contexts than all other age groups. While aggression was most probable in *In kennel towards people* and *Interactions with unfamiliar people* contexts for dogs aged 4 months through 6 years, dogs over 6 years old were most likely to show aggression in the *Handling* context, leading to interactions between, for example, *Handling* and *In kennel towards people*, and between *Handling* and *Interactions with unfamiliar people*

513 contexts compared to the other age groups (Fig 1b; Table S3). Aggression when *Eating*
514 *food* and in *Interactions with toys* contexts also increased compared to that expressed
515 by younger dogs, resulting in credible differences between, for instance, *Eating food* and
516 *Interactions with familiar people* contexts between dogs aged 10 months to 3 years and
517 over 6 years (OR = 0.379; CI: 0.300 to 0.465) and between *Out of kennel towards people*
518 and *Interactions with toys* contexts between over 6 year olds and all other age groups
519 (Table S3).

520

a



b

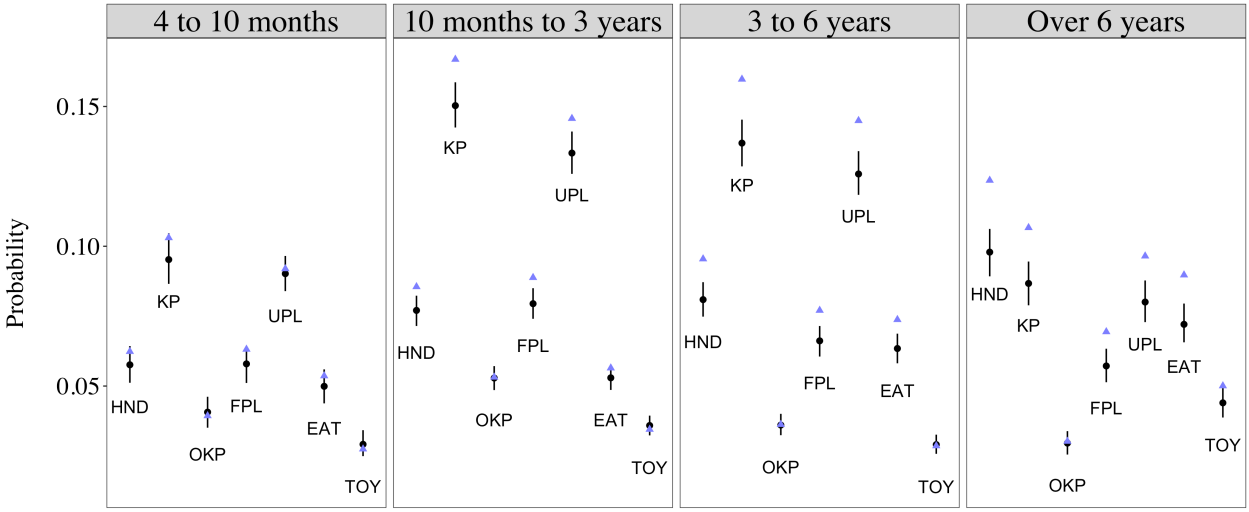


Fig 1. Predicted probabilities of aggression towards people in different contexts by sex (panel a) and age groups (panel b). Black points and vertical lines show mean and 95% credibility intervals of model parameter estimates; blue triangles show raw sample data. Abbreviations used in the figure: HND (*Handling*); KP (*In kennel towards people*); OKP (*Out of kennel towards people*); FPL (*Interactions with familiar people*); UPL (*Interactions with unfamiliar people*); EAT (*Eating food*); TOY (*Interactions with toys*).

521

522

Aggressiveness towards dogs

The odds of aggression towards dogs, across categorical predictors and for an average dog of mean weight and length of stay at the shelter, was 0.176 (CI: 0.168, 0.184), corresponding to a probability of approximately 15%. Dogs were most likely to show aggression in the *Interactions with male dogs* context (OR = 0.297; CI: 0.198 to 0.217) and least likely in the *In kennel towards dogs* context (OR = 0.099; CI: 0.094 to 0.104; Fig 2; Table S4).

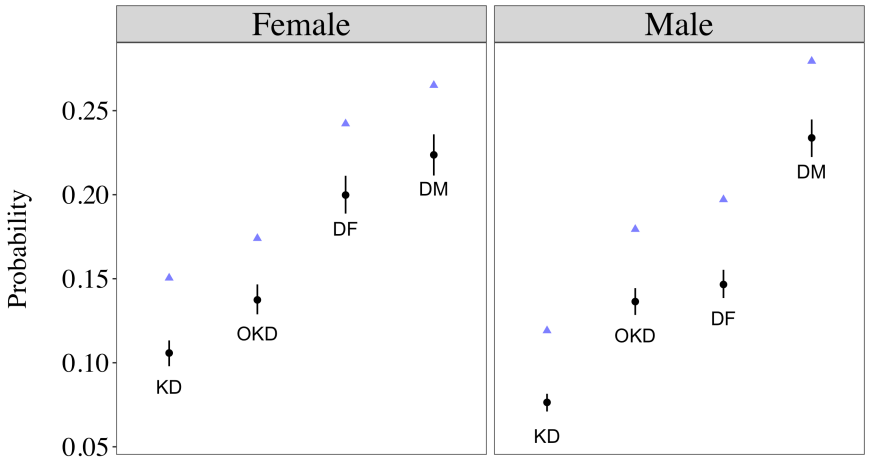
No credible mean-level differences existed between females and males (OR = 1.187; CI: 1.128 to 1.250). However, the difference in aggression between the *Interactions with female dogs* and *Interactions with male dogs* contexts was smaller for females (OR = 1.542; CI: 1.400 to 1.704; Fig 2a; Table S4), as were the differences between *Interactions with male dogs* and *In kennel towards dogs* (OR = 0.661; CI: 0.590 to 0.732) and *In kennel towards dogs* and *Out of kennel towards dogs* (OR = 1.420; CI: 1.269 to 1.587). Females were also more likely to show aggression in *Interactions with female dogs* than *Out of kennel towards dogs* compared to males (OR = 1.444; CI: 1.301 to 1.603).

Dogs aged 4 to 10 months old had credibly lower odds of aggression towards dogs than older dogs across contexts (Fig 2b; Table S4). However, contexts and age also showed interactive effects. In particular, aggression in *Interactions with female dogs* and

543 *Interactions with male dogs* contexts tended to increase relative to other contexts. For
544 instance, the relationship between *Interactions with female dogs* and *Out of kennel*
545 *towards dogs* contexts reversed in direction between 4 to 10 month and 10 month to 3
546 year olds (OR = 0.595; CI: 0.495 to 0.688) as did the relationship between *Interactions*
547 *with male dogs* and *Out of kennel towards dogs* contexts (OR = 0.499; CI: 0.422 to
548 0.575). The relationship between *In kennel towards dogs* and *Out of kennel towards*
549 *dogs* contexts also changed across age groups (Fig 2b; Table S4). Four to 10 months old
550 were more likely to show aggression in *Out of kennel towards dogs* than *In kennel*
551 *towards dogs* contexts, but the difference was smaller in 10 months to 3 year olds (OR =
552 0.608; CI: 0.505 to 0.728) and in over 6 year olds (OR = 0.396; CI: 0.316 to 0.481). The
553 latter relationship was reversed in 3 to 6 year olds compared to 4 to 10 month old dogs
554 (OR = 0.277; CI: 0.227 to 0.331) and 10 month to 3 year old dogs (OR = 0.456; CI: 0.396
555 to 0.516).

556

a



b

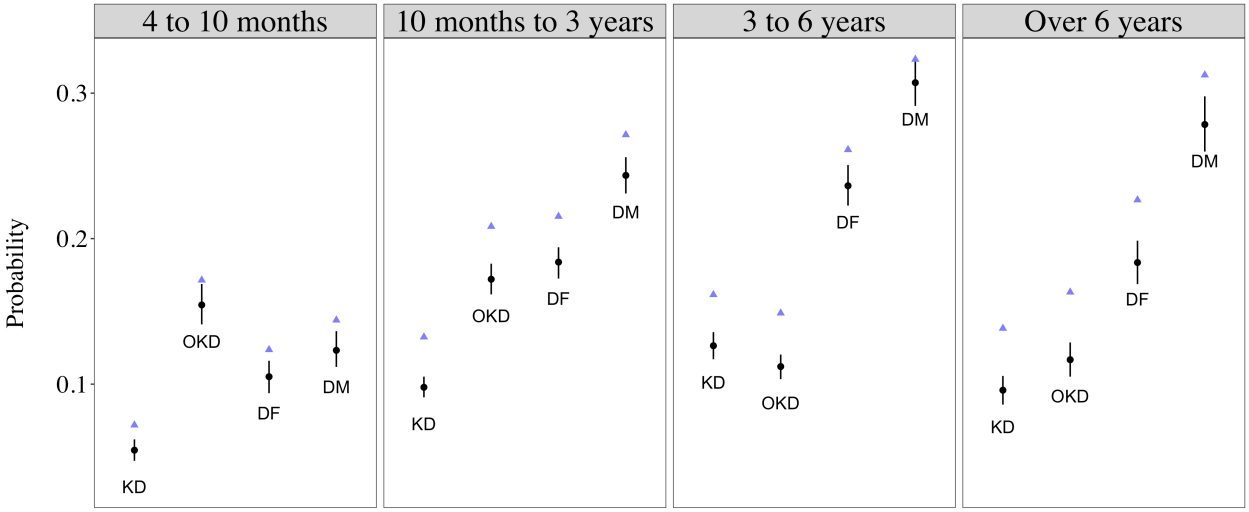


Fig 2. Predicted probabilities of aggression towards dogs in different contexts by sex (panel a) and age groups (panel b). Black points and vertical lines show mean and 95% credibility intervals of model parameter estimates; blue triangles show raw sample data. Abbreviations used in the figure: KD (*In kennel towards dogs*); OKD (*Out of kennel towards dogs*); DF (*Interactions with female dogs*); DM (*Interactions with male dogs*).

557

558

Repeatability

Both aggressiveness towards people and dog showed moderate repeatability across contexts (ICC_{people} = 0.479; CI: 0.466, 0.491; ICC_{dogs} = 0.303; CI: 0.291, 0.315), although aggressiveness towards people was more repeatable than aggressiveness towards dogs ($ICC_{difference}$ = 0.176; CI: 0.158, 0.192).

Discussion

In this study, we have examined whether local independence and measurement invariance hold for hypothesised latent aggressiveness traits in shelter dogs. Observational recordings of aggression directed towards people and dogs across different shelter contexts were explained by two positively correlated latent variables, and behaviour across contexts was moderately repeatable. These results are consistent with the definition of animal personality as behaviour that shows moderately consistent between-individual differences across time or contexts [4], and characterised by multiple observed behaviours being decomposed into lower-dimensional behavioural traits. Yet, subsequent investigations indicated violations of local independence and measurement invariance, questioning the validity of the latent variables as homogeneous personality traits. While a number of factors may contribute to the low predictive validity of certain dog personality assessments [48], ensuring the robustness of inferences made about personality traits is critical, especially considering the large

number of traits that have been proposed [34,35]. Given the popularity of latent variable models, such as exploratory and confirmatory factor analysis (e.g. [17,29,71]), to understand the organisation of personality in dogs, ascertaining local independence and measurement invariance of personality traits should be routine practice, as it has become in human (personality) psychology (e.g. [26,72–74]).

Local independence implies that the association between manifest variables is greater than that explained by the latent variable. Here, local independence was investigated between manifest variables that were believed to have close temporal-spatial relationships. While local independence was confirmed for contexts reflecting aggressiveness towards dogs, contexts reflecting aggressiveness towards people showed a violation of local independence. Aggression in the *Handling* context shared a negative residual correlation with both the *In kennel towards people* and *Interactions with unfamiliar people* contexts, while positive residual correlations were present between *Handling* and *Interactions with toys*, and *Out of kennel towards people* and *Interactions with unfamiliar people* contexts (Table S4). Violations of local independence may arise through shared method variance [75–78] or unmodelled latent variables influencing manifest variables [79,80]. For example, if a dog showed aggression when an unfamiliar person approached, it may be less likely to be handled by that person, inducing a negative residual correlation conditional on latent levels of aggression as was observed here between the *Handling* and *Interactions with unfamiliar people* contexts. Likewise, the *Handling* context and the *Interactions with toys* contexts are similar in that both

601 require dog and person to be close together, and interacting with toys and handling may
602 co-occur at the same time, inducing a positive residual correlation between contexts.
603
604 While authors have argued that greater standardisation and validation of personality
605 assessments is key to ensuring the accurate measurement of underlying traits [34,46], it
606 may be untenable to avoid dependencies between testing contexts. Displays of
607 aggression in one sub-test will likely change how people conduct future sub-tests with
608 the same dog, regardless of test standardisation. Moreover, the hierarchical structure of
609 animal personality, including the presence of behavioural syndromes, makes the
610 isolated measurement of one trait unrealistic [75]. For instance, the positive residual
611 correlation between *Out of kennel towards people* and *Interactions with unfamiliar*
612 *people* may be mediated by additional traits of interest to personality researchers, such
613 as fearfulness or anxiety [29,81], if dogs who are fearful of interacting with unfamiliar
614 people are more likely to show aggression beyond that described by a latent
615 aggressiveness trait. Some human psychologists have argued that violations of local
616 independence are a natural consequence of the organisation of behaviour as a complex
617 dynamic system [82,83], which unfolds with respect to time- and context-dependent
618 constraints [84]. Thus, awareness of local independence and its violation could facilitate
619 closer understanding of the dynamics driving personality test responses beyond
620 explanations purely based on personality traits.
621

While different subsets of a population may differ in mean levels of trait expression, interactions between behavioural responses and those subsets indicate that the same phenomenon is not under measurement across groups [23,24]. Using a Bayesian hierarchical model analogous to the Rasch model, we found that the probability of aggression across contexts was still dependent on sex and age conditional on latent levels of aggressiveness towards people and dog (Figs 1 and 2; Tables S3 and S4), indicating the violation of measurement invariance. Female dogs, for example, were more likely than males to show aggression in *Out of kennel towards people* and *Interactions with unfamiliar people* contexts relative to other contexts (Fig 1a). Females also demonstrated similar odds of aggression during *Interactions with female dogs* and *Interactions with male dogs*, whereas males were more likely to show aggression towards male than female dogs (Fig 2a). Thus, latent levels of aggressiveness did not easily explain differences in aggression across contexts. As with local independence, different behavioural variables unaccounted for in this study may also result in violations of measurement invariance. For instance, while dogs up to 6 years old were most likely to show aggression in *In kennel towards people* and *Interactions with unfamiliar people contexts*, dogs over 6 years old demonstrated aggression most commonly in the *Handling* context, which may reflect an increase in pain-motivated aggression. Dogs over 6 years old also showed an increase in aggression in the *Eating food* and *Interactions with toys* contexts relative to other age groups, suggesting that older dogs in shelter populations may be less tolerant during close interactions with

643 people (i.e. handling, people in the vicinity of their food and toys) compared to other
644 contexts.

645

646 Investigating factors that predict a dog's personality, or whether a dog's personality
647 predicts other outcomes of interest (e.g. future behavioural scores; [39,40]), is also of
648 substantive interest to researchers. Persson *et al.* [32] found interactive effects between
649 sex and age on human-directed social behaviour traits. Moreover, Asp *et al.* [56] found
650 that sex and age interacted with breed in explaining differences in C-BARQ trait scores.
651 To ensure the robustness of their conclusions, however, researchers should ensure that
652 the traits being investigated satisfy model assumptions, not only for statistical accuracy,
653 but so that interpreting the difference between individuals as a function of trait scores is
654 both feasible and meaningful.

655

656 Local independence and measurement invariance are assumptions of confirmatory,
657 reflective latent variable models, but are not required assumptions for formative
658 models, which posit that the latent variable is simply a linear composite of manifest
659 variables, rather than a causal, underlying variable [20]. While formative models such as
660 principal components analysis (e.g. [32,33,85]) may, as a result, appear attractive and
661 continue to be used in dog personality studies, their use has been discouraged. Principal
662 components analysis will always result in lower-dimensional variables comprised of
663 linear combinations of manifest variables, even when those manifest variables are

uncorrelated random variables (e.g. see [12]). Consequently, finding principal components that underlie behavioural data is neither surprising nor evidence for the discovery of domain-general personality traits. Crucially, for personality traits to be of use in understanding the organisation of dog behaviour or be considered as predictors of future dog behaviour, they should, arguably, hold causal status. In human psychology, Schimmack [86] and Borsboom [6] note the importance of interpreting latent variables as causal variables for understanding unobserved constructs, such as personality traits. To this end, reflective models, especially confirmatory approaches such as structural equation modelling or item response theory, present a more powerful framework to distinguish signal from noise in multivariate behavioural data [20,86] and are concurrent with a theoretical interpretation of personality traits as causal variables underpinning animal behaviour [16]. Increasing the popularity of such approaches could be particularly helpful in evaluating the reproducibility of dog personality traits across existing studies.

Although we have identified violations of both local independence and measurement invariance, we remain cautious about hypothesising *a posteriori* about their causes. A problem for the wider perspective of animal personality research is that personality traits are typically defined operationally, based on the statistical repeatability of quantifiable behaviour [77,87,88]. As has been discussed in human personality psychology, operational definitions are ontologically ambiguous [89,90]. That is, while operational definitions facilitate experimentation in animal personality [4], they are

686 ambiguous with respect to the biological mechanisms underlying trait expression. For
687 example, Budaev and Brown remark that boldness, defined as a propensity to take risks,
688 could encompass a range of distinct personality traits, each with a different biological
689 basis [75]. Whilst reflective latent variable models allow researchers to test hypotheses
690 about the relatedness of measured behaviours via one or more underlying traits, they
691 have also been criticised as ambiguous [81]. For example, it is uncertain what reflective
692 latent variables may represent in biological organisation [89] or even whether they are
693 features individuals possess or simply emergent features of between-individual
694 differences [91,92]. Such considerations highlight the importance of research on the
695 proximate mechanisms of personality [87] and longitudinal data analyses to separate
696 between- from within-individual behavioural variation [93,94].

697

698 A number of authors have emphasised the poor predictive value of aggression tests in
699 shelter dogs compared to tests of other traits [37,39,48]. The low occurrence of
700 aggression can make its accurate measurement difficult [39], and some studies actively
701 exclude dogs that have shown aggression in the shelter (e.g. [39]). The probability of
702 observing aggression recorded in this study was low, especially in contexts reflecting
703 aggressiveness towards people (Fig 1). Nonetheless, evaluation of the validity of the
704 behavioural recordings indicated that shelter employees might mistake observations of
705 aggression for non- aggressive responses (e.g. over-excitement and frustration when
706 seeing people/dogs), meaning that the true probability of aggression was potentially
707 under-estimated (although incorrectly coding other behaviours as aggression also

occurred, albeit rarely). Infrequent occurrence and/or recording of aggression may also limit accurate predictions of future behaviour. Patronek and Bradley [48] demonstrate using simulation that the low prevalence of aggression inflates the chance that aggression shown in a shelter assessment represents a false positive. In general, our results support this conclusion in the sense that aggression may be shown differentially across contexts not explained by latent levels of aggressiveness. Violations of local independence and measurement invariance as found here indicate, further, that it is not only the difference between false and true positives and negatives, but the validity of inferring homogeneous personality traits by which to compare individual dogs, that needs careful consideration. Consequently, we agree with recommendations to establish the efficacy of longitudinal, observational assessments rather than relying on a single assessment made using a traditional test battery [31,39,48]. This approach will prioritise the cumulative understanding of a dog's context-dependent behaviour and help to guide decisions about the potential risk a dog poses to humans and other animals.

Conclusion

This study has tested the assumptions of local independence and measurement invariance of personality traits in shelter dogs. Using structural equation modelling, aggression across behavioural contexts was explained by two, correlated latent variables and demonstrated repeatability. Nevertheless, significant residual correlations

729 remained between certain behavioural contexts related to aggressiveness towards
730 people, violating the assumption of local independence. In addition, aggression in
731 different contexts showed differential patterns of response across sex and age,
732 indicating a lack of measurement invariance. Violations of local independence and
733 measurement invariance imply that aggressiveness towards people and dogs may not
734 be enough to explain patterns of aggression in different contexts, or that inferences
735 based on these hypothesised personality traits may in fact be misleading. We encourage
736 researchers to more closely assess the measurement assumptions underlying reflective
737 latent variable models before making conclusions about the effects of, or factors
738 influencing, personality.

739

740 Acknowledgements

741 The authors are grateful to Battersea Dogs and Cats Home for allowing us to access the
742 data for this study.

743

744 References

745 1. Dochtermann NA, Jenkins SH. Behavioural syndromes in Merriam's kangaroo rats
746 (*Dipodomys merriami*): a test of competing hypotheses. Pro R Soc Biol Sci. 2007;274:
747 2343–2349. doi:10.1098/rspb.2007.0622

2. Dochtermann NA, Dingemanse NJ. Behavioral syndromes as evolutionary constraints. Behav Ecol. 2013; art002. doi:10.1093/beheco/art002

3. Westneat DF, Wright J, Dingemanse NJ. The biology hidden inside residual within-individual phenotypic variation. Biol Rev Camb Philos Soc. 2015;90: 729–743. doi:10.1111/brv.12131

4. Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ. Integrating animal temperament within ecology and evolution. Biol Rev Camb Philos Soc. 2007;82: 291–318. doi:10.1111/j.1469-185X.2007.00010.x

5. Sih A, Bell AM, Johnson JC, Ziemba RE. Behavioral syndromes: an integrative overview. Q Rev Biol. 2004;79: 241–277. doi:10.1086/422893

6. Spearman C. “General intelligence,” objectively determined and measured. Am J Psychol. 1904;15: 201–292. doi:10.2307/1412107

7. Beaujean AA. Latent variable modeling using R: A step-by-step guide. New York: Routledge; 2014.

8. Bollen KA. Latent variables in psychology and the social sciences. Annu Rev Psychol. 2002;53: 605–634. doi:10.1146/annurev.psych.53.100901.135239

9. Borsboom D. The attack of the psychometricians. Psychometrika. 2006;71: 425–440. doi:10.1007/s11336-006-1447-6

766 10. Budaev DS. How many dimensions are needed to describe temperament in animals:
767 a factor reanalysis of two data sets. Int J Comp Psychol. http://cogprints.org/5478/;
768 1998.

769 11. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation
770 perspective. Psychol Bull. 1991;110: 305–314. doi:10.1037/0033-2909.110.2.305

771 12. Budaev SV. Using principal components and factor analysis in animal behaviour
772 research: caveats and guidelines. Ethology. 2010;116: 472–480. doi:10.1111/j.1439-
773 0310.2010.01758.x

774 13. Dingemanse NJ, Dochtermann N, Wright J. A method for exploring the structure of
775 behavioural syndromes to allow formal comparison within and between data sets. Anim
776 Behav. 2010;79: 439–450. doi:10.1016/j.anbehav.2009.11.024

777 14. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of
778 exploratory factor analysis in psychological research. Psychol Methods. 1999;4: 272–
779 299. doi:10.1037/1082-989X.4.3.272

780 15. Preacher KJ, MacCallum RC. Repairing Tom Swift's electric factor analysis machine.
781 Understanding Statistics. 2003;2: 13–43. doi:10.1207/S15328031US0201_02

782 16. Araya-Ajoy YG, Dingemanse NJ. Characterizing behavioural "characters": An
783 evolutionary framework. Proc R Soc B. 2014;281: 20132645.
784 doi:10.1098/rspb.2013.2645

785 17. Arden R, Adams MJ. A general intelligence factor in dogs. Intelligence. 2016;55: 79–
786 85. doi:10.1016/j.intell.2016.01.008

787 18. van den Berg SM, Heuven HCM, van den Berg L, Duffy DL, Serpell JA. Evaluation of
788 the C-BARQ as a measure of stranger-directed aggression in three common dog breeds.
789 Appl Anim Behav Sc. 2010;124: 136–141. doi:10.1016/j.applanim.2010.02.005

790 19. Bartholomew DJ. Factor analysis for categorical data. J R Stat Soc Series B Stat
791 Methodol. 1980;42: 293–321.

792 20. Markus KA, Borsboom D. Frontiers of test validity theory: Measurement, causation,
793 and meaning. 1st ed. New York, N.Y: Routledge; 2013.

794 21. Drasgow F. Study of the measurement bias of two standardized psychological tests.
795 Journal of Applied Psychology. 1987;72: 19–29. doi:10.1037/0021-9010.72.1.19

796 22. Mellenbergh GJ. Item bias and item response theory. Int J Educ Res. 1989;13: 127–
797 143. doi:10.1016/0883-0355(89)90002-5

798 23. Meredith W. Measurement invariance, factor analysis and factorial invariance.
799 Psychometrika. 1993;58: 525–543. doi:10.1007/BF02294825

800 24. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response
801 theory: two approaches for exploring measurement invariance. Psychol Bull. 1993;114:
802 552–566. doi:10.1037/0033-2909.114.3.552

803 25. Jansen BRJ, van der Maas HLJ. Statistical test of the rule assessment methodology by
804 latent class analysis. Dev Rev. 1997;17: 321–357. doi:10.1006/drev.1997.0437

805 26. Wicherts JM, Dolan CV. Measurement invariance in confirmatory factor analysis: an
806 illustration using IQ test performance of minorities. Educ Meas Issues Pract. 2010;29:
807 39–47. doi:10.1111/j.1745-3992.2010.00182.x

808 27. Widaman KF, Reise SP. Exploring the measurement invariance of psychological
809 instruments: applications in the substance use domain. In: Bryant KJ, Windle M, West
810 SG, editors. The science of prevention: Methodological advances from alcohol and
811 substance abuse research. Washington, DC, US: American Psychological Association;
812 1997. pp. 281–324.

813 28. Goddard ME, Beilharz RG. Genetic and environmental factors affecting the suitability
814 of dogs as Guide Dogs for the Blind. Theor Appl Genet. 1982;62: 97–102.
815 doi:10.1007/BF00293339

816 29. Hsu Y, Serpell JA. Development and validation of a questionnaire for measuring
817 behavior and temperament traits in pet dogs. J Am Vet Med Assoc. 2003;223: 1293–
818 1300. doi:10.2460/javma.2003.223.1293

819 30. Scott JP, Fuller JL. Genetics and the social behavior of the dog. University of Chicago
820 Press; 2012.

821 31. Rayment DJ, Groef BD, Peters RA, Marston LC. Applied personality assessment in
822 domestic dogs: limitations and caveats. Appl Anim Behav Sci. 2015;163: 1–18.
823 doi:10.1016/j.applanim.2014.11.020

824 32. Persson ME, Wright D, Roth LSV, Batakis P, Jensen P. Genomic regions associated
825 with interspecies communication in dogs contain genes related to human social
826 disorders. Sci Rep. 2016;6. doi:10.1038/srep33439

827 33. Sundman A-S, Johnsson M, Wright D, Jensen P. Similar recent selection criteria
828 associated with different behavioural effects in two dog breeds. Genes Brain Behav.
829 2016;15: 750–756. doi:10.1111/gbb.12317

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

833 35. Fratkin JL, Sinn DL, Patall EA, Gosling SD. Personality consistency in dogs: a meta-
834 analysis. PLOS ONE. 2013;8: e54907. doi:10.1371/journal.pone.0054907

835 36. Posluns JA, Anderson RE, Walsh CJ. Comparing two canine personality assessments:
836 convergence of the MCPQ-R and DPQ and consensus between dog owners and dog
837 walkers. Appl Anim Behav Sci. doi:10.1016/j.applanim.2016.12.013

838 37. Bennett SL, Litster A, Weng H-Y, Walker SL, Luescher AU. Investigating behavior
839 assessment instruments to predict aggression in dogs. Appl Anim Behav Sci. 2012;141:
840 139–148. doi:10.1016/j.applanim.2012.08.005

841 38. Mohan-Gibbons H, Weiss E, Slater M. Preliminary investigation of food guarding
842 behavior in shelter dogs in the United States. Animals. 2012;2: 331–346.
843 doi:10.3390/ani2030331

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

848 40. Riemer S, Müller C, Virányi Z, Huber L, Range F. The predictive value of early
849 behavioural assessments in pet dogs: a longitudinal study from neonates to adults. PLOS
850 ONE. 2014;9: e101237. doi:10.1371/journal.pone.0101237

851 41. Wilsson E, Sundgren P-E. Behaviour test for eight-week old puppies: heritabilities of
852 tested behaviour traits and its correspondence to later behaviour. Appl Anim Behav Sci.
853 1998;58: 151–162. doi:10.1016/S0168-1591(97)00093-2

854 42. Goodloe LP, Borchelt PL. Companion dog temperament traits. J Appl Anim Welf Sci.
855 1998;1: 303–338. doi:10.1207/s15327604jaws0104_1

856 43. Jones AC. Development and validation of a dog personality questionnaire. Doctoral
857 Thesis. University of Texas at Austin. 2008. Available from:
858 http://gosling.psy.utexas.edu/wp-content/uploads/2014/10/Amanda-Claire-Jones-Diss-
859 2008.pdf

860 44. Orritt R. Dog bites: a complex public health issue. Vet Rec. 2015;176: 640–641.
861 doi:10.1136/vr.h3215

862 45. Svartberg K. Shyness-boldness predicts performance in working dogs. Appl Anim
863 Behav Sci. 2002;79: 157–174. doi:10.1016/S0168-1591(02)00120-X

864 46. Taylor KD, Mills DS. The development and assessment of temperament tests for
865 adult companion dogs. J Vet Behav. 2006;1: 94–108. doi:10.1016/j.jveb.2006.09.002

866 47. Henning G. Meanings and implications of the principle of local independence. Lang
867 Res. 1989;6: 95–108. doi:10.1177/026553228900600108

868 48. Patronek GJ, Bradley J. No better than flipping a coin: reconsidering canine behavior
869 evaluations in animal shelters. J Vet Behav. 2016;15: 66–77.
870 doi:10.1016/j.jveb.2016.08.001

871 49. Casey RA, Loftus B, Bolster C, Richards GJ, Blackwell EJ. Human directed aggression
872 in domestic dogs (*Canis familiaris*): occurrence in different contexts and risk factors.
873 Appl Anim Behav Sci. 2014;152: 52–63. doi:10.1016/j.applanim.2013.12.003

874 50. Hsu Y, Sun L. Factors associated with aggressive responses in pet dogs. Appl Anim
875 Behav Sci. 2010;123: 108–123. doi:10.1016/j.applanim.2010.01.013

876 51. Sherman CK, Reisner IR, Taliaferro LA, Houpt KA. Characteristics, treatment, and
877 outcome of 99 cases of aggression between dogs. Appl Anim Behav Sci. 1996;47: 91–
878 108. doi:10.1016/0168-1591(95)01013-0

52. Voith VL, Trevejo R, Dowling-Guyer S, Chadik C, Marder A, Johnson V, et al. Comparison of visual and DNA breed identification of dogs and inter-observer reliability. Sociology. 2013;3: 17–29.

53. Owczarczak-Garstecka, SC, Burman OHP. Can sleep and resting behaviour be used as indicators of welfare in shelter dogs (*Canis lupus familiaris*)? PLOS ONE. 2016;11: e0163620. doi:10.1371/journal.pone.0163620

54. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. https://www.r-project.org/

55. Honaker J, King G, Blackwell M. Amelia: A program for missing data. 2015. Version 1.7.4. https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

56. Asp HE, Fikse WF, Nilsson K, Strandberg E. Breed differences in everyday behaviour of dogs. Appl Anim Behav Sci. 2015;169: 69–77. doi:10.1016/j.applanim.2015.04.010

57. Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., et al, Lavaan: Latent Variable Analysis. 2016. Version 0.5-22. https://cran.r-project.org/web/packages/lavaan/

58. Muthén B, Christoffersson A. Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika. 1981;46: 407–419. doi:10.1007/BF02293798

59. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49: 115–132. doi:10.1007/BF02294210

60. Jorgensen TD, Pornprasertmanit S, Miller P, Schoemann A, Rosseel Y, Quick C, et al. semTools: useful tools for structural equation modeling. 2016. https://rdrr.io/cran/semTools/

61. Hermida R. The problem of allowing correlated errors in structural equation modeling: concerns and considerations. Comp Method Soc Sci. 2015;3: 05–17.

62. Millsap RE, Yun-Tein J. Assessing factorial invariance in ordered-categorical measures. Multivariate Behav Res. 2004;39: 479–515. doi:10.1207/S15327906MBR3903_4

63. Kamata A. Item analysis by the hierarchical generalized linear model. J Educ Meas. 2001;38: 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x

64. Van den Noortgate W, De Boeck P. Assessing and explaining differential item functioning using logistic mixed models. J Educ Behav Stat. 2005;30: 443–464.

65. Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev Camb Philos Sci. 2010;85: 935–956. doi:10.1111/j.1469-185X.2010.00141.x

66. Stan Development Team. Stan modeling language users guide and reference manual. 2016. Version 2.14.0. https://github.com/stan-dev/stan/releases/download/v2.14.0/stan-reference-2.14.0.pdf

67. Kruschke JK. Bayesian data analysis. Wiley Interdiscip Rev: Cogn Sci. 2010;1: 658–676. doi:10.1002/wcs.72

68. Vehtari A, Gelman A, Gabry J, Piironen J, Goodrich B. Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. 2016. Version 1.0.0. https://cran.r-project.org/web/packages/loo/loo.pdf

69. Kruschke J. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. 2nd ed. Academic Press; 2014.

70. Chen JJ, Tsong Y, Kang S-H. Tests for equivalence or noninferiority between two proportions. Drug Inf J. 2000;34: 569–578. doi:10.1177/009286150003400225

71. Barnard S, Marshall-Pescini S, Passalacqua C, Beghelli V, Capra A, Normando S, et al. Does subjective rating reflect behavioural coding? Personality in 2 month-old dog puppies: an open-field test and adjective-based questionnaire. PLOS ONE. 2016;11: e0149831. doi:10.1371/journal.pone.0149831

72. Ion A, Iliescu D. The measurement equivalence of personality measures across high- and low-stake test taking settings. Pers Individ Dif. 2017;110: 1–6. doi:10.1016/j.paid.2017.01.008

935 73. Bowden SC, Saklofske DH, van de Vijver FJR, Sudarshan NJ, Eysenck SBG. Cross-
936 cultural measurement invariance of the Eysenck Personality Questionnaire across 33
937 countries. Pers Individ Dif. 2016;103: 53–60. doi:10.1016/j.paid.2016.04.028

938 74. van den Berg SM, Moor MHM de, McGue M, Pettersson E, Terracciano A, Verweij
939 KJH, et al. Harmonization of Neuroticism and Extraversion phenotypes across
940 inventories and cohorts in the Genetics of Personality Consortium: an application of
941 item response theory. Behav Genet. 2014;44: 295–313. doi:10.1007/s10519-014-9654-x

942 75. Budaev S, Brown C. Personality traits and behaviour. In: Brown C, Laland K, Krause J,
943 editors. Fish cognition and behavior. Wiley-Blackwell; 2011. pp. 135–165.

944 76. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-
945 multimethod matrix. Psychol Bull. 1959;56: 81–105. doi:10.1037/h0046016

946 77. Carter AJ, Feeney WE, Marshall HH, Cowlishaw G, Heinsohn R. Animal personality:
947 what are behavioural ecologists measuring? Biol Rev Camb Philos Soc. 2013;88: 465–
948 475. doi:10.1111/brv.12007

949 78. Podsakoff PM, MacKenzie SB, Lee J-Y, Podsakoff NP. Common method biases in
950 behavioral research: a critical review of the literature and recommended remedies. J
951 Appl Psychol. 2003;88: 879–903. doi:10.1037/0021-9010.88.5.879

952 79. Cole DA, Ciesla JA, Steiger JH. The insidious effects of failing to include design-driven
953 correlated residuals in latent-variable covariance structure analysis. Psychol Methods.
954 2007;12: 381–398. doi:10.1037/1082-989X.12.4.381

80. Yen WM. Scaling performance assessments: strategies for managing local item dependence. J Educ Meas. 1993;30: 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

81. Carter AJ, Marshall HH, Heinsohn R, Cowlishaw G. How not to measure boldness: novel object and antipredator responses are not the same in wild baboons. Anim Behav. 2012;84: 603–609. doi:10.1016/j.anbehav.2012.06.015

82. Cramer AOJ, van der Sluis S, Noordhof A, Wichers M, Geschwind N, Aggen SH, et al. Dimensions of normal personality as networks in search of equilibrium: you can't like parties if you don't like people. Eur J Pers. 2012;26: 414–431. doi:10.1002/per.1866

83. Schmittmann VD, Cramer AOJ, Waldorp LJ, Epskamp S, Kievit RA, Borsboom D. Deconstructing the construct: a network perspective on psychological phenomena. New Ideas Psychol. 2013;31: 43–53. doi:10.1016/j.newideapsych.2011.02.007

84. Hamaker EL, Nesselroade JR, Molenaar PCM. The integrated trait-state model. J Res Pers. 2007;41: 295–315. doi:10.1016/j.jrp.2006.04.003

85. Fadel FR, Driscoll P, Pilot M, Wright H, Zulch H, Mills D. Differences in trait impulsivity indicate diversification of dog breeds into working and show lines. Sci Rep. 2016;6. doi:10.1038/srep22162

86. Schimmack U. What multi-method data tell us about construct validity. Eur J Pers. 2010;24: 241–257. doi:10.1002/per.771

87. Duckworth RA. Neuroendocrine mechanisms underlying behavioral stability: implications for the evolutionary origin of personality. Ann N Y Acad Sci. 2015;1360: 54–74. doi:10.1111/nyas.12797

88. Koski SE. How to measure animal personality and why does It matter? Integrating the psychological and biological approaches to animal personality. In: Inoue-Murayama M, Kawamura S, Weiss A, editors. From Genes to Animal Behavior. Springer Japan; 2011. pp. 115–136. doi:10.1007/978-4-431-53892-9_5

89. D. Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge University Press; 2005.

90. Maul A, Torres Irribarra D, Wilson M. On the philosophical foundations of psychological measurement. Measurement. 2016;79: 311–320. doi:10.1016/j.measurement.2015.11.001

91. Molenaar PCM. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Measurement. 2004;2: 201–218. doi:10.1207/s15366359mea0204_1

92. Adolf J, Schuurman NK, Borkenau P, Borsboom D, Dolan CV. Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. Front Psychol. 2014;5. doi:10.3389/fpsyg.2014.00883

992 93. Stamps JA, Briffa M, Biro PA. Unpredictable animals: individual differences in
993 intraindividual variability (IIV). Anim Behav. 2012;83: 1325–1334.
994 doi:10.1016/j.anbehav.2012.02.017

995 94. Biro PA, Stamps JA. Using repeatability to study physiological and behavioural traits:
996 ignore time-related change at your peril. Anim Behav. 2015;105: 223–230.
997 doi:10.1016/j.anbehav.2015.04.008

998

999

Supporting Information

Table S1. Counts of aggression per context. The number of dogs who had 0, 1, and > 1 counts of aggression.

Table S2. Tetrachoric correlations between aggression contexts. Tetrachoric correlations between aggression contexts on the raw binary data, before the multiple imputation. Abbreviations used: HND (*Handling*); FPL (*Interactions with familiar people*); UPL (*Interactions with unfamiliar people*); KD (*In kennel towards dogs*); KP (*In kennel towards people*); OKD (*Out of kennel towards dogs*); OKP (*Out of kennel towards people*); EAT (*Eating food*); TOY (*Interactions with toys*); DM (*Interactions with male dogs*); DF (*Interactions with female dogs*).

Table S3. Bayesian hierarchical model parameter estimates for aggression towards people in different contexts. Mean and 95% credibility interval (CI) estimates for all parameters from the Bayesian hierarchical logistic model assessing measurement invariance for contexts reflecting aggressiveness towards people. Differences between levels of categorical variables are indicated by '.v.' in the parameter name; interactions are denoted with '*' in the parameter name. The decision rule for each parameter is given except for those variables not interpreted inferentially: YES = 95% CI falls

completely outside the region of practical equivalence (ROPE); NULL = 95% CI falls completely inside the ROPE; ROPE = 95% CI partly covers the ROPE.

Table S4. Bayesian hierarchical model parameter estimates for aggression towards dogs in different contexts. Mean and 95% credibility interval (CI) estimates for all parameters from the Bayesian hierarchical logistic model assessing measurement invariance for contexts reflecting aggressiveness towards dogs. Differences between levels of categorical variables are indicated by '.v.' in the parameter name; interactions are denoted with '*' in the parameter name. The decision rule for each parameter is given except for those variables not interpreted inferentially: YES = 95% CI falls completely outside the region of practical equivalence (ROPE); NULL = 95% CI falls completely inside the ROPE; ROPE = 95% CI partly covers the ROPE.