Assessing personality traits in dogs: conceptual and methodological issues

Evaluering av personalighestrekk hos hund: konseptuelle og metodiske aspekter

Philosophiae Doctor (PhD) Thesis

Conor Goold

Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural Sciences

Ås 2017

Thesis number 2017:72
ISSN 1894-6402
ISBN 978-82-575-1468-6

Supervisors

Prof. Ruth C. Newberry
Department of Animal and Aquacultural Sciences
Faculty of Biosciences
Norwegian University of Life Sciences
PO Box 5003, 1432 Ås
Norway

Dr. Judit Vas
Department of Animal and Aquacultural Sciences
Faculty of Biosciences
Norwegian University of Life Sciences
PO Box 5003, 1432 Ås
Norway

Prof. Bjarne O. Braastad
Department of Animal and Aquacultural Sciences
Faculty of Biosciences
Norwegian University of Life Sciences
PO Box 5003, 1432 Ås
Norway

Acknowledgements

Too many people have helped me through these last three years. I wrote and re-wrote this section a few times, but couldn't do the acknowledgements justice on paper, so I'll just keep it short and sweet.

Firstly, my main supervisor Ruth is a superstar. Her support and encouragement throughout has been more than any PhD student or friend could wish for. Secondly, my co-supervisors Judit and Bjarne have been incredibly patient and supportive of my work and me as a student, happy for me to explore my ideas independently and be there to provide feedback whenever I needed it. Thirdly, to everyone else in the Ethology research group: thank you for being so welcoming and putting up with my terrible Norwegian (or considerable lack of) for so long. Thirdly, a huge thanks goes to Peter Laurie, Ali Taylor, Nathalie Ingham, and the canine behaviourists at Battersea Dogs and Cats Home for allowing and aiding me in gathering the data for two of the papers in this thesis. Last, but not least, to my family and closest friends. You know who you are; I couldn't have done it without you.

Contents

1 List of papers **vii**

2 Summary **ix**

3 Sammendrag **xiii**

4 Introduction **1**

4.1 Animal personality: concepts and conundrums 1

4.2 Two approaches to studying animal personality 2

4.2.1 The operational approach 2

4.2.2 The latent variable approach 6

4.3 In need of a third approach? The network perspective 8

4.3.1 Methodological concerns with the operational and latent variable approaches 8

4.3.2 The network perspective 10

5 Dog personality **13**

5.1 Personality traits in dogs 13

5.2 Personality consistency in dogs 15

5.3 Approaches to animal personality: where do dogs sit? 17

6 Aims of the thesis **21**

7 Materials & Methods **23**

7.1 Shelter dogs 23

7.1.1 Observational behavioural assessment 24

7.2 Police dogs 25

7.2.1 Questionnaire 25

7.3 Validity & inter-rater reliability 26

7.4 Data analysis 27

7.4.1 Validity & inter-rater reliability 27

7.4.2 Missing data 27

7.4.3 Inferential models 27

7.5 Ethical approval 30

7.6 Data accessibility 30

8 Results and Discussion **31**

8.1 Assumptions of latent variable approaches 31

8.2 Beyond personality 33

8.3 Applying the network perspective 35

8.4 Limitations & future directions 36

9 Conclusion **39**

References **41**

10 Papers in order **53**

1 List of papers

Paper I

Goold C., Newberry RC. (2017). Aggressiveness as a latent personality trait of domestic dogs: testing local independence and measurement invariance. Accepted in *PLoS ONE*.

Paper II

Goold C., Newberry RC. (2017). Modelling personality, plasticity and predictability in shelter dogs. Accepted with minor revisions in *Royal Society Open Science*.

Paper III

Goold C., Vas J., Olsen C., Newberry RC. (2016). Using network analysis to study behavioural phenotypes: an example using domestic dogs. *Royal Society Open Science*. 3:160268.

2 Summary

Animal personality is defined by consistent between-individual differences in behaviour through time or across contexts. Behaviour is further organised into broader behavioural dimensions referred to as personality *traits* (e.g. fearfulness, aggressiveness or boldness). While animal personality is a relatively new field, researchers have been interested in quantifying and predicting stable behavioural traits or dimensions in domestic dogs (*Canis lupus familiaris*) for over fifty years. Nonetheless, deciding which personality traits are most relevant or which traits behaviours reflect remains a difficult task for animal (as well as human) personality researchers. Largely, this is because personality is something we infer from behavioural data rather than directly observe, which depends on the conceptual and methodological approach taken. For dogs in particular, the predictive validity of personality assessments has been inconsistent, such as in behavioural assessments of shelter dogs. Moreover, there have been a diverse number of traits and behavioural dimensions proposed, with little consensus across studies on which traits are most relevant for describing dog behaviour.

This thesis evaluated conceptual and methodological issues of assessing personality and personality traits in dogs. In particular, the papers addressed key aspects of the statistical analysis of behavioural data on dogs for making inferences about personality and personality traits, drawing upon perspectives across both ethology and human psychology. The papers demonstrate three broad results.

First, research to understand which personality traits underlie dog behaviour would benefit from moving from largely exploratory-based to hypothesis-driven frameworks. Personality traits in dogs are usually inferred by using exploratory latent variable statistical models, such as principal components analysis, and studies have applied a mixture of latent variable models that have differing underlying assumptions. Confirmatory, reflective latent variable models provide a more powerful framework for testing competing hypotheses about the latent structure of behavioural data in dogs and for verifying the robustness of the derived personality traits. Using data on inter-context aggressive behaviour towards people and dogs in shelter dogs, we found two, correlated latent variables: aggressiveness towards people and dogs, respectively. However, these posited traits failed to account for all of the co-variation between aggressive behaviour across contexts, violating the assumption of *local independence*. Moreover, interactions between aggression contexts and the sex and age of the dogs demonstrated a violation of *measurement invariance*. That is, sex and age differences in aggressive behaviour could not be simply explained by differences in latent aggressiveness traits. The robustness and reproducibility of other personality traits in dogs could be verified by applying similar approaches to multivariate data.

Secondly, dogs do not only differ in how they behave on average (i.e. personality), but in the amount they change their behaviour across time (*behavioural plasticity*) and the amount of day-to-day fluctuation around their average behaviour (*predictability*). By applying the framework of behavioural reaction norms, popular within behavioural and evolutionary ecology, we studied these different components of variation in dogs' reactions to meeting unfamiliar people over time at a shelter. Accounting for individual differences in intra-individual behaviour (i.e. plasticity and predictability) in addition to personality improved the predictive accuracy of our results compared to focusing on personality variation only. The results also highlighted the importance of gathering repeated measurements on individuals when estimating behavioural variation. Specifically, behavioural predictions at the individual level were highly uncertain compared to those at the group-level (aggregating data across dogs), since the amount of data available on each dog individually was often small. Together, these results emphasised the benefits of longitudinal assessments of dog behaviour in shelters, and the importance of systematic modelling of both inter-individual (i.e. personality) and intra-individual variation in dog behaviour.

Thirdly, predominant approaches to conceptualising of animal personality traits are faced with a number of challenges. Inspired by recent work in human psychology, we elucidated how animal personality, and integrated behavioural phenotypes in general, can be re-conceptualised using a network perspective. The network perspective represents the behavioural repertoire of individuals as a system of causally connected, autonomous behaviours. Behavioural dimensions or traits are, thus, viewed as emergent patterns of causally related clusters of behaviours, rather than separate underlying variables. We demonstrated the application of network analysis to survey data collected on behavioural and motivational characteristics of police patrol and detection dogs. Our analyses emphasised a number of close, functional relationships between variables consistent with previous research on dog personality, as well as unique insights from novel network statistics into the organisation of police dog behaviour. We highlighted the merits of this perspective for furthering work on the organisation of behavioural phenotypes and animal personality, and situating this research within work on a diverse range of complex systems across science.

In summary, this thesis has drawn upon advancements across ethology and human psychology to present novel directions for understanding personality in dogs. The work will be of benefit to researchers determining which personality traits explain individual differences in dog behaviour and those aiming to predict future dog behaviour. Lastly, the results should stimulate a greater awareness of the conceptual issues involved in making inferences about personality in dogs and other animals.

3 Sammendrag

Dyrs personlighet er definert som konsistente forskjeller i atferd mellom individer over tid eller på tvers av ulike sammenhenger, kontekster. Atferden er videre organisert i bredere atferdsdimensjoner som kalles personlighetstrekk (for eksempel fryktsomhet, aggressivitet eller dristighet). Selv om dyrs personlighet er et relativt nytt felt, har forskere vært interessert i å kvantifisere og forutsi stabile atferdsegenskaper eller atferdsdimensjoner hos hunder (*Canis lupus familiaris*) i over femti år. Likevel, å avgjøre hvilke personlighetstrekk som er mest relevante eller hvilke egenskaper en atferd reflekterer, er fortsatt en vanskelig oppgave for personlighetsforskere på dyr (og mennesker). Stort sett skyldes dette at personlighet er noe vi analyserer utfra atferdsdata i stedet for å observere direkte, og noe som avhenger av den konseptuelle og metodologiske tilnærmingen som er gjort. For spesielt hunder har personlighetsvurderinger ikke gitt konsekvente forutsigelser av hundens atferd, for eksempel i bedømmelser av atferd hos hunder i omplasseringsinstitusjoner (hjelpesentre). Videre har det vært foreslått varierende antall atferdstrekk og atferdsdimensjoner, med liten konsensus på tvers av studier angående hvilke trekk som er mest relevante for å beskrive hundens atferd.

Denne doktoravhandlingen evaluerte konseptuelle og metodiske aspekter i forbindelse med vurdering av personlighet og personlighetstrekk hos hunder. Artiklene behandlet viktige aspekter ved den statistiske analysen av atferdsdata fra hunder for å beskrive personlighet og personlighetstrekk, og de benytter perspektiver på tvers av etologi og humanpsykologi. Artiklene viser tre brede resultater.

For det første, forskning for å forstå hvilke personlighetstrekk som ligger til grunn for hundens atferd vil ha nytte av å endres fra et hovedsakelig undersøkelsesbasert til et hypotesebasert utgangspunkt. Personlighetstrekk hos hunder er vanligvis utledet ved å bruke statistiske modeller med utforskende latente variable, for eksempel prinsipalkomponentanalyse, og studier har benyttet en blanding av modeller med latente variabler som har ulike underliggende forutsetninger. Bekreftende, reflekterende modeller med latente variabler gir et kraftigere rammeverk for å teste konkurrerende hypoteser om den latente strukturen av atferdsdata hos hunder, og slike modeller kan verifisere robustheten av de utledede personlighetstrekkene. Ved å bruke data om aggressiv atferd i ulike sammenhenger rettet mot mennesker og hunder i omplasseringsinstitusjoner, fant vi to korrelerte latente variabler: aggressivitet mot henholdsvis mennesker og hunder. Disse egenskapene forklarte imidlertid ikke all

samvariasjon mellom aggressiv atferd på tvers av sammenhenger, noe som er i strid med antagelsen om lokal uavhengighet. Videre viste interaksjoner mellom aggresjonskontekster og kjønn og alder hos hundene et brudd på prinsippet om måleinvariasjon. Det vil si at forskjeller i aggressiv atferd med hensyn på kjønn og alder ikke kunne forklares bare av forskjeller i latente aggressivitetstrekk. Robustheten og reproduserbarheten av andre personlighetstrekk hos hunder kunne bekreftes ved å anvende liknende tilnærminger til multivariate data.

For det andre varierer hundene ikke bare i hvordan de oppfører seg i gjennomsnitt (dvs. personligheten), men i hvor mye de endrer sin atferd over tid (atferdsplastisitet) og i hvor store svingninger det er fra dag til dag i forhold til den gjennomsnittlige atferden (forutsigbarhet). Ved å ta utgangspunkt i atferdsreaksjonsnormer, som er populært innen atferdsøkologi og evolusjonær økologi, studerte vi disse forskjellige variasjonskomponentene i hunders reaksjoner når de møter ukjente mennesker over tid i et omplasseringssenter. Ved å ta hensyn til individuelle forskjeller i intraindividuell atferd (dvs. plastisitet og forutsigbarhet) i tillegg til personlighet, kunne vi forbedre nøyaktigheten i forutsigelsene av resultatene våre sammenliknet med når vi fokuserer kun på personlighetsvariasjon. Resultatene fremhevet også betydningen av å foreta gjentatte målinger på enkeltindivider ved estimering av atferdsvariasjon. Spesielt var atferdsprediksjoner på individnivå svært usikre sammenliknet med dem på gruppenivå (samlet for alle hundene), siden datamengden som var tilgjengelig for hver hund ofte var for liten. Sammen understreket disse resultatene fordelene ved langsgående vurderinger av hundens atferd i omplasseringsinstitusjonene, og betydningen av systematisk modellering av variasjoner i hundens atferd både innen individet (dvs. personlighet) og mellom individer.

For det tredje møter de mest vanlige tilnærmingene til konseptualisering av dyrs personlighetstrekk en rekke utfordringer. Inspirert av nylige arbeider innen humanpsykologi belyste vi hvordan dyrs personlighet, og integrerte atferdsfenotyper generelt, kan konseptualiseres på nytt ved hjelp av et nettverksperspektiv. Nettverksperspektivet består i å analysere individets atferdsrepertoar som et system med kausalt forbundne, autonome atferder. Atferdsdimensjoner eller atferdstrekk betraktes således som fremvoksende mønstre av kausalt relaterte atferdsklynger, i stedet for separate underliggende variabler. Vi demonstrerte anvendelsen av nettverksanalyse for å undersøke data fra skjemaer for atferdstrekk og motivasjonstrekk hos politiets patrulje- og søkshunder. Våre analyser understreket en rekke tette, funksjonelle relasjoner mellom variabler som er i tråd med tidligere undersøkelser av hunders personlighet, samt unik innsikt ervervet fra ny nettverksstatistikk om hvordan politihunders atferd er organisert. Vi fremhevet fordelene ved dette perspektivet for å fremme arbeid med organisering av atferdsfenotyper og dyrs personlighet, og plassere denne forskningen innen arbeid på et mangfold av ulike komplekse systemer på tvers av vitenskaper.

For å summere opp, denne avhandlingen har dratt nytte av fremskritt innen etologi og humanpsykologi for å presentere nye retninger for å forstå personlighet hos hunder. Arbeidet vil være til nytte for forskere som vil avklare hvilke personlighetstrekk som forklarer individuelle forskjeller i hunders atferd og for de som har som mål å forutsi fremtidig atferd hos hunder. Til slutt bør resultatene stimulere til en større bevissthet om de konseptuelle problemene som er involvert når en skal lage utledninger om personlighet hos hunder og andre dyr.

4 Introduction

> *"When observers spend hours recording behaviour, they end up not only with behavioural data, but clear impressions of individuals."*
>
> Stevenson-Hinde *et al.* (1980)

4.1 Animal personality: concepts and conundrums

Understanding individual differences in humans has been of scientific interest for over a century (Spearman, 1904), and individuality is central to numerous discussions in modern society. Although the importance of variation among non-human animals has been recognised since Charles Darwin outlined the theory of evolution by natural selection, individual differences in animals has notably become of scientific interest in the previous twenty years, a topic most generally referred to as *animal personality*. Animal personality is now relevant to a range of topics in animal behaviour, including cognition (Carere and Locurto, 2011), behavioural and evolutionary ecology (Réale *et al.*, 2007), experimental biology (Roche *et al.*, 2016), and applied animal behaviour (Gosling and John, 1999; Rayment *et al.*, 2015). Personality has, further, been studied in a variety of taxa, including fish, amphibians, insects, birds and mammals (Bell *et al.*, 2009).

Personality is a term that is familiar to everyone, but is much harder to define and investigate scientifically. The seminal definition of personality in humans is *those characteristics of individuals that describe and account for consistent patterns of feeling, thinking and behaving* (Pervin and John, 1999), with particular 'characteristics' being defined as personality traits (McCrae and Costa, 1995). Animal behaviourists define personality more narrowly as *consistent between-individual differences in behaviour across time or contexts* (Réale *et al.*, 2007). In both fields, the exact terminology used to refer to personality and the relative scientific merits of personality research have been variable. In human psychology, personality has been considered as, on one hand, an integral biological basis determining human behaviour (e.g. McCrae and Costa, 1995) whilst, on the other hand, neither a biological nor psychological category but purely an "ethical and spiritual category" (Hutton, 1945, p. 165). In animals, while the difficulties with studying personality are accepted, many authors have highlighted merits of studying personality (Réale *et al.*, 2007; Gosling and John, 1999; Briffa and Weiss, 2010), although in some areas scientists argue that personality adds little to existing theories or is unncessarily anthropomorphic (Crews, 2013; DiRienzo and Montiglio, 2015; Beekman and Jordan, 2017). Crews (2013, p. 875) writes, for example, that "this new anthropomorphism [i.e. personality] is unnecessary and should be viewed with skepticism".

Understanding and researching animal personality is difficult, in part, due to inconsistencies in the theoretical foundations and goals of personality research (David and Dall, 2016; Uher, 2011). This instantiates in a variety of methods used to quantify personality (Carter *et al.*, 2013; Koski, 2011), as well as a diverse collection of terminology to describe personality traits (e.g. 'characteristics', 'dimensions', 'characters'; Jones and Gosling, 2005; Uher, 2011). Therefore, before turning to the main focus of this thesis, which is the topic of assessing personality in domestic dogs (*Canis lupus familiaris*), it is beneficial to outline some general approaches to investigating animal personality.

4.2 Two approaches to studying animal personality

Animal personality studies could be categorised in a number of ways, from using Tinbergen's Four Questions to identify which perspectives of personality any one study is investigating (i.e. the development, causation, functional value and/or evolution of personality; e.g. Dall *et al.*, 2004), to the data collection methods used to learn about individuals' personalities (e.g. subjective assessments, or behavioural codings in observational data and/or experiments; Carter *et al.*, 2013). Authors have also employed broader 'meta-thereotical' categorisations, which are useful because they encapsulate the more specific, downstream methodological decisions taken when studying personality. Two approaches have been distinguished in the latter categorisations, which I refer to here as the *operational* and *latent variable* approaches.

4.2.1 The operational approach

To avoid making connotations with psychological dispositions, Réale *et al.* (2007) provided an operational definition of animal *temperament* as behaviour that differs consistently between individuals through time and across contexts, which has since been adopted as the definition of animal personality. This definition of animal personality is sufficiently general to pertain to any quantifiable behaviour believed to reflect a personality trait of interest (i.e. the behaviour should be ecologically relevant; Carter *et al.*, 2013), and "does not make any assumptions about either the underlying proximate mechanisms for personality variation or what types of behavior should be considered personality traits" (Duckworth, 2015, p. 2). At the same time, Réale *et al.* (2007) proposed five main traits or axes of animal temperament: shyness-boldness, exploration-avoidance, activity, aggressiveness, and sociability. Since the focus of this operational definition is on quantifying between-individual variation in one or more measured behaviours through time or across contexts, it has inherited the statistical frameworks of quantitative genetics, namely hierarchical regression models (Dingemanse and Dochtermann, 2013).

Imagine an experiment where one measures the activity behaviour of dog puppies placed in an unfamiliar room once in each of two experimental conditions: when the owner is present and, subsequently, when the owner is absent. The behaviour recorded could be the number of times a puppy crosses a set of marked grid lines on the floor (McGarrity *et al.*, 2015). Each puppy in the experiment will have two behavioural recordings, one from each condition (owner present and owner absent). A hierarchical linear regression model analysing the outcome variable (y) at instance i for each dog j could be written as:

$$y_{ij} = \alpha + \mu_j + \beta E_i + \epsilon_i \tag{1}$$

The terms in this model include i) a y-intercept parameter (α), denoting the overall mean number of times dogs crossed the marked floor lines, ii) an individual dog-specific intercept parameter (μ_j), describing the deviation from α for the particular dog j, a coefficient (β) describing how much the mean activity behaviour changes depending on if the particular observation was recorded in the 'owner present' ($E = 1$) or 'owner absent' ($E = 2$) conditions, and finally a residual error term (ϵ_i) capturing the difference between the expected value from the model and the actual recorded value. Across all dogs, the vector of μ_j parameters are taken to be normally distributed with mean zero and standard deviation σ_μ, written as $\mu_j \sim N(0, \sigma_\mu)$. Similarly, the residual error terms are assumed distributed $\epsilon_i \sim N(0, \sigma_\epsilon)$. By mean-centering the covariate E, the intercept parameter and the individual dog-specific intercepts are calculated at the 'average' environment. A visual representation of this model is presented in Figure 2 *(a)*, where activity behaviour is seen to be lower in the owner absent condition compared to the owner present condition (i.e. a negative slope).

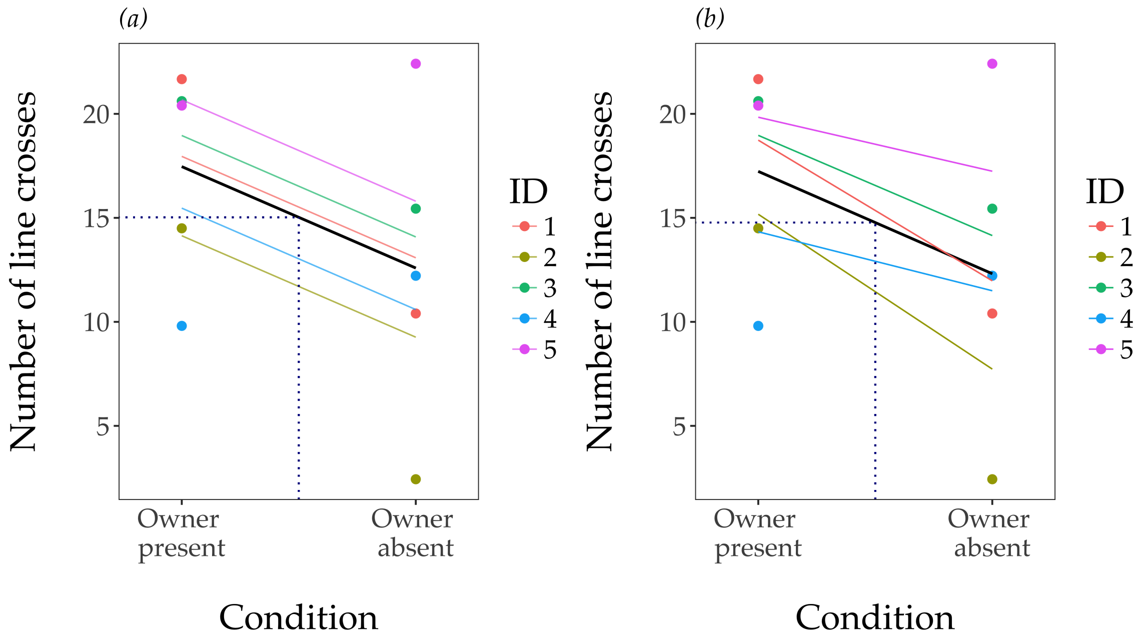


Figure 1: The statistical operationalisation of personality using hierarchical regression models. Imagine recording the number of times dog puppies ($N = 5$, here) cross a set of lines marked on the floor on two consecutive occassions, once when the owner is present and once when the owner is absent. Hierarchical linear regression is used to analyse the data: (a) the black line shows the average regression line estimated across dogs, and the intercept parameters are allowed to vary by dog as deviations from the average (evaluated across conditions denoted by the dotted blue line); (b) the same as (a) but with the slope parameter also varying by dog.

The most important summary metric of the operational approach is behavioural repeatability, defined as the proportion of total variance explained by between-individual differences (Nakagawa and Schielzeth, 2010). Repeatability is calculated using the intraclass correlation coefficient (ICC). For the model above, this is:

$$ICC = \frac{\sigma_{\mu}^2}{\sigma_{\mu}^2 + \sigma_{\epsilon}^2} \tag{2}$$

The ICC is usually around 40% in studies of animal personality (Bell *et al.*, 2009). In some cases, such as when there are only two repeated measurements of behaviour, repeatability is inferred from other types of correlation coefficients, such as the widely-known Pearson's product moment correlation. However, correlation coefficients such as Pearson's correlation (or Spearman's rank-order correlation coefficient for non-Gaussian data) are typically measures of relative consistency (i.e. how consistent individuals are relative to other individuals) rather a reflection of the absolute agreement of scores for any one individual through time (Nakagawa and Schielzeth, 2010). Nonetheless, the interpretation of the ICC can also change depending upon its specific calculation (McGraw and Wong, 1996). For example, partialling out systematic influences of time (e.g. including day or week of measurement in the statistical model) means the ICC reflects relative consistency more than absolute agreement (see Biro and Stamps, 2015 for a discussion of ignoring time in estimates of repeatability).

The above statistical formulation completes the operationalisation of personality. In this example, the intercepts quantify between-individual differences in activity behaviour across contexts and the ICC reflects behavioural repeatability. The power of this modelling framework is in its flexibility for understanding other aspects of behavioural variation. For example, note that activity behaviour does not change the same way for all dogs in Figure 2 *(a)*: some individuals become more active when the owner is absent (e.g. individual 5). Thus, the assumption that each dog's behaviour should be modelled with the same slope parameter is likely too stringent. Figure 2 *(b)* shows the result of varying the slope parameter by dog as well as the intercept, known as *behavioural plasticity* (Dingemanse *et al.*, 2010). Since there are only five dogs and each dog only has two observations, the individual-specific slopes are still largely influenced by the group-level negative slope (e.g. individual 5 has a more positive, but still negative, slope), a statistical property known as hierarchical shrinkage or partial pooling. Nonetheless, individual differences in the slope parameters are still evident and better represent the data. Hierarchical regression models can be extended further to take into account individual differences in the intra-individual residual variance between dogs, which has been called *behavioural predictability*, or to include non-linear functions across time or contexts.

Together, this formulation has become known *behavioural reaction norms* (Dingemanse *et al.*, 2010; Cleasby *et al.*, 2015), akin to the use of reaction norms to study phenotypic plasticity in evolutionary biology more generally (e.g. Nussey *et al.*, 2007). I refer to this approach as operational because personality, as well as plasticity and predictability, are inferred purely with reference to how the behaviour is measured and subsequently analysed (Bridgman, 1954; see also Borsboom (2005) for a synthesis of operational definitions of psychological constructs). In contrast, Koski (2011) referred to how personality is studied in behavioural and evolutionary ecology as the 'biological' approach, while Carter *et al.* (2013) described it as a reductive approach. Moreover, the operational approach most similarly reflects the individual-oriented approaches discussed by Uher (2011) because the emphasis on longitudinal modelling of behaviour allows disentangling between- from within-individual variation (e.g. personality versus plasticity). However, what are these approaches being distinguished from?

4.2.2 The latent variable approach

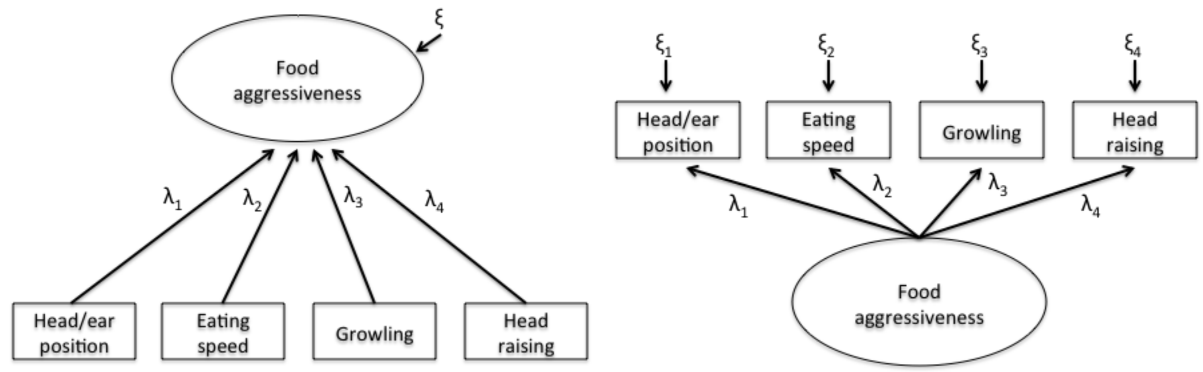
The second main approach to studying animal personality considered here is termed the latent variable approach. Latent variable approaches focus on discovering underlying or *latent* variables explaining covariance between a number of measured variables (the *manifest* variables). Unlike the operational approach, personality traits are conceptualised as superordinate, biological variables to be inferred from behavioural data, rather than operationally-defined constructs. Because latent personality traits are not directly observed, the broad class of latent variable statistical models are highly popular in both human and animal personality research to infer which personality traits or dimensions explain behaviour (Bollen and Lennox, 1991). Indeed, latent variable methods have been popular in human psychology for over a century (Spearman, 1904). Koski (2011) referred to this latent variable approach as the 'psychological' approach to studying animal personality, partly because it is often applied to survey data completed by people knowledgeable about individual animals in applied animal behaviour, similar to self-report methods in human psychology. Nonetheless, latent variable methods are increasing in popularity among behavioural and evolutionary ecologists also (e.g. Araya-Ajoy and Dingemanse, 2014; Dochtermann and Jenkins, 2007; Martin and Suarez, 2017), making this distinction unclear. Latent variable approaches also resonate with the variable-oriented perspective considered by Uher (2011), since the primary goal is to understand which personality traits can be related to measured patterns of behaviour at the population level, rather than concerted modelling of between- and within-individual differences.

The advantage of latent variable models is the ability to specify and estimate the relationship between the measured, manifest variables and the latent, scientific constructs of interest. Two varieties of models are available: *formative* and *reflective* (Beaujean, 2014). Formative models assume that the latent variables are simple linear composites of the manifest variables (causal indicator models; Bollen and Lennox, 1991). Principal components analysis is a formative model, recommended when a multivariate data set requires reducing into a smaller number of variables that retain most of the (co)variation in the data. As such, principal components analysis will always return components, even when the manifest variables are uncorrelated, random variables (Budaev, 2010). By contrast, reflective models assume that the manifest variables are caused by the latent variables, with some degree of measurement error (effect indicator models; Bollen and Lennox, 1991). Consequently, there are not simply a data reduction tool, but a powerful measurement model estimating the causal relationship between a number of observed and unobserved variables. Reflective models can either be exploratory (e.g. exploratory factor analysis) or confirmatory (e.g. confirmatory factor analysis, structural equation modelling), with the latter providing flexibility in testing and comparing *a priori* hypotheses via metrics of model fit (Beaujean, 2014).

Consider an experiment studying food aggressiveness in dog puppies, where a puppy is given a bowl of food and, subsequently, an experimenter attempts to remove the bowl with a fake, plastic hand for safety. Imagine we record four different behavioural variables during the experiment on ordinal scales: i) ear and tail position (e.g. relaxed to tense), ii) eating speed, iii) the amount of growling, and iv) the amount of head raising (as discussed by McGarrity *et al.*, 2015). We can visually represent both formative and reflective latent variable models for this example as path diagrams (Figure 2).

Choosing a formative or reflective model is dependent on the substantive research question and the overall goal of the analysis. While there may be cases in which a formative model may be more appropriate for studying personality traits (see the next section) or even cases where the difference between them is small for practical purposes (Velicer and Jackson, 1990), there is consensus across the human and animal personality literatures that reflective models are most suitable given that personality traits are often permitted causal status on the expression of behaviour (e.g. humans: Fabrigar *et al.*, 1999; Preacher and MacCallum, 2003; Borsboom, 2006; animals: Budaev, 2010; Araya-Ajoy and Dingemanse, 2014). In the previous example, a change in the level of food aggressiveness would be expected to result in a change in the recorded behaviours, and the recorded behaviours would be expected to correlate with each other because they all reflect the same construct (McGarrity *et al.*, 2015). This points towards a reflective model, where food aggressiveness is not simply a composite variable for separate unrelated behaviours, but an underlying dimension that influences the expression of these recorded behaviours. Formative models have much weaker assumptions about the manifest variables, which do not need to be correlated with each other or show any internal consistency (Bollen and Lennox, 1991), criteria that are usually considered necessary for discovering personality traits in behavioural data (Taylor and Mills, 2006; Carter *et al.*, 2013).

Despite their differences, it is still the case that formative and reflective models are used interchangeably in animal personality (Budaev, 2010). Similarly, while distinctions between different types of latent variable models are more readily discussed in human psychology, psychometricians have warned against an over-reliance on forma-



(a) Formative model

(b) Reflective model

Figure 2: Path diagrams for two types of latent variable model: formative and reflective. Formative models assume that the latent variable (denoted as a circle) is a linear composite of the manifest variables (denoted as squares), while reflective models assume that the latent variable causally influences the manifest variables. The coefficients estimating the relationship between the latent and manifest variables are denoted λ, while error variances are denoted with ξ. Here, food aggressiveness (the latent construct) is measured by four different manifest variables.

tive instead of reflective latent variable models for studying psychological constructs (Borsboom, 2006).

4.3 In need of a third approach? The network perspective

4.3.1 Methodological concerns with the operational and latent variable approaches

The operational and latent variable approaches are both powerful ways of studying personality and individual differences in animals, but also possess a number of conceptual and methodological shortcomings. The operational approach makes the experimental analysis of personality easier by operationalising measured behaviours as personality traits, but appears to lack the theoretical foundations to provide a rigorous framework for studying personality. For example, Dochtermann and Nelson (2014) found that two operational measurements of exploration in house crickets (*Acheta domesticus*) showed consistent between-individual differences in behaviour. However, the measurements were uncorrelated with each other, contrary to their predictions if both behaviours were in fact reflections of exploration. As the authors note, it is difficult to understand these findings using a purely operational definition of exploration, and they highlight that too little attention has been placed on the conceptual basis of animal personality traits. In fact, operationalism as a philosophy of science (Bridgman, 1954) and operational definitions of psychological constructs in human psychology (e.g. classical test theory; Borsboom, 2005; Maul *et al.*, 2016) have receieved much criticism also (e.g. Green, 1992). As Maul *et al.* (2016) summarise, "theoretical concepts are seldom exhausted by their operational definitions", and Borsboom (2005) notes that operational definitions are ontologically ambiguous. Indeed, although Réale *et al.* (2007) suggested an operational definition of temperament in animals to avoid making connotations to underlying dispositions, they invoke similar concepts when writing "we assume that the behaviour of the mouse in an open field reveals its reactions to a new and open environment and thus its *exploratory tendencies*" (Réale *et al.*, 2007, p. 304; emphasis added). The difference between explaining behaviour by alluding to 'tendencies' rather than 'dispositions' appears trivial, and places operational definitions on uncertain ground.

Latent variable approaches possess the advantage that they explictly model the relationship between observed and unobserved variables. Yet, problems arise when there is not enough scientific theory to warrant such formal modelling. Notably, it is rarely the case that the posited latent variable can be identified in biological organisation. For instance, although latent variable models have been used for over a century to define intelligence in humans (Spearman, 1904), sometimes known as the *g* factor, no biological referent has been identified (van der Maas *et al.*, 2014; van der Maas *et al.*, 2006). Is this necessarily a problem? A number of authors believe that for latent variables to be of real use as scientific constructs in human psychology, a position of scientific realism is necessary (Borsboom, 2005; Schimmack, 2010; Anusic and Schimmack, 2016). That is, there is a need to interpret the latent variables causally for using them to make predictions about behaviour, or in discovering predictors of variation in the latent variables. For example, it is difficult to study the ontogeny of sex differences in intelligence when the latent intelligence variable is not interpreted as a real, causal entity.

Reflective latent variable models, further, have a number of assumptions that may be unrealistic. The assumption of *local independence*, for instance, states that the latent variable accounts for the correlations between the manifest variables (Markus and Borsboom, 2013; Epskamp *et al.*, 2016b). That is, since reflective models assume that the latent variable causes variation in the manifest variables, the manifest variables should be independent conditional on the latent variable. Another important assumption is *measurement invariance* (Reise *et al.*, 1993; Markus and Borsboom, 2013; Wicherts and Dolan, 2010), which is satisfied when the structural relationships between the latent variable and manifest variables are maintained in different subsets of the population (e.g. within individuals, age groups or sex). Consider the reflective latent variable example of food aggressiveness in the preceding section and shown in Figure 2 *(b)*. Imagine we first fit this model across a large population of dogs, and then fit the same model for male and female dogs separately. While males and females may differ in their average levels of food aggressiveness (e.g. males may have higher levels of food aggressiveness than females), measurement invariance asserts that the estimated parameters (e.g. the λ coefficients) are the same. If they are not the same, any differences between males and females cannot be simply attributed to differences in food aggressiveness itself, because the measurement relationship is different. While local independence and measurement invariance may be too strict in many cases (Markus and Borsboom, 2013), they are amenable to verifaction in the modelling process, meaning researchers can empirically assess the suitability of a reflective latent variable model more easily than the suitability of a formative model.

4.3.2 The network perspective

An emerging approach in human psychology is the network perspective (Cramer *et al.*, 2012; Schmittmann *et al.*, 2013). A network is a system of components that interact with each other in dynamic ways, and can be represented as a graphical model where the components are typically denoted as *nodes* and the relationships between the components as *edges*. A correlation network of the food aggressiveness behaviours discussed in the preceding section is shown in Figure 3. Network analysis has been used to model a wide range of complex dynamic systems across science (Kolaczyk and Csárdi, 2014), including neuroscience, ecology and evolution, and animal behaviour (e.g. brain networks: Bullmore and Sporns, 2009; physiological regulatory networks: Cohen *et al.*, 2012; ecological networks: Proulx *et al.*, 2005; animal social networks: Croft *et al.*, 2008).

The network perspective in psychology posits that behavioural, cognitive and affective components form correlated dimensions because those components possess causal relationships with each other. One of the largest applications of network analysis has been to a range of psychopathological disorders, such as major depression disorder (Cramer *et al.*, 2016). While a latent variable approach envisages a set of symptoms being caused by the same underlying disorder, a network approach suggests that the disorder emerges when the symptoms form a causally connected unit. Lack of sleep and problems with concentration are two symptoms of major depression, and are expected to have causal relationships (i.e. lack of sleep causes problems with concentration the next day, and potentially vice versa), even in non-depressed individuals.

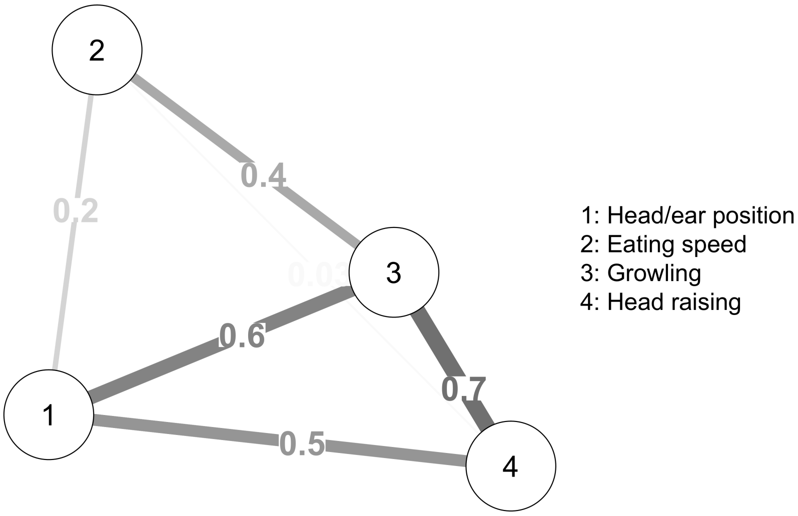


Figure 3: Network of food aggressiveness behaviours (nodes) and their positive correlations (edges; numbers represent Pearson correlation coefficients). Rather than these behaviours being the cause of, or simply formulating, a latent food aggressiveness variable (as shown in Figure 2), the network perspective would envisage food aggressiveness as an emergent property of the direct, causal relationships between these behaviours.

But when those symptoms become causally connected to, and temporally dependent on, other symptoms (e.g. feelings of worry, loss of appetite), the individual slips into a depressed state (van Borkulo *et al.*, 2015). The network perspective has also been applied to personality psychology, such as aspects of the Five Factor model (e.g. Schmittmann *et al.*, 2013) and intelligence (van der Maas *et al.*, 2006).

Causality in this instance is defined in terms of conditional independence relationships, following the work of Pearl (2009). That is, given a set of correlated variables (behavioural, cognitive or affective components) believed to be associated with a certain construct, we can hypothesise a causal relationship between two variables when they remain correlated after partitioning out the effects of the remaining variables. In a network, these relationships are expressed as partial correlations, and many advances have been made in recent years on the estimation of regularised graphical models in psychology (e.g. Gaussian graphical models; Epskamp *et al.*, 2016b).

Psychological constructs, such as personality traits, in the network perspective are *emergent properties* of the causal relationships between cognitive, affective and behavioural components. Simply stated, an emergent property of a complex system is one that only exists when parts of a system assemble together and one that is more than the sum of its parts (Kauffman, 1993; Capra and Luisi, 2014; Bar-Yam, 2016). Thus, food aggressiveness cannot be reduced to just one of the food aggressiveness behaviours (Figure 3), but requires the presence of all the behaviours acting in concert.

Conceptualsing personality traits as emergent properties is, in fact, most similar to a formative modelling approach (e.g. van der Maas *et al.*, 2014; Schmittmann *et al.*, 2013), where the components are considered to be relatively autonomous and coalesce to form a higher-order variable (Bollen and Lennox, 1991). However, there are some important differences.

The same emergent property of a complex dynamic system may arise through different causal pathways, a phenomenom known as *degeneracy* (Edelman and Gally, 2001; Seifert *et al.*, 2016). For personality, this means that the same 'traits' or functional network structure can emerge despite individual differences in the actual connections between components. For example, two dogs described as 'food aggressive' may display each behaviour at differing intensities, and the pattern of causal relationships between the behaviours for each dog (i.e. individual-specific networks; Bringmann *et al.*, 2013) may not be the same. Network analysis, further, offers a number of unique ways of quantifying the structure of complex systems. One metric is node *centrality*, a family of statistics that identify nodes which are important for maintaining network structure. *Betweenness* centrality, for example, measures the number of shortest paths between all nodes that run through each node (Brandes, 2001). Nodes that have higher betweenness centrality are, thus, expected to have greater influence on the behaviour of other nodes in the network. These insights and the flexibility offered by a network perspective, and complex systems theory more generally, cannot be accrued from a formative modelling approach.

In summary, the network approach provides a different way to conceptualise the multidimensional organisation of the behavioural phenotype that is concommitant with many other areas of science studying complex systems. Consequently, adopting a network perspective may advance the clarity of how personality and personality traits are defined and studied.

5 Dog personality

While animal personality is a relatively new field, researchers have been interested in quantifying individual differences and behavioural traits in domestic dogs for half a century (e.g. see Scott and Fuller, 2012 for a summary of many early experiments). Now, the field of dog personality encompasses research on selecting the best service or working dogs (Goddard and Beilharz, 1982; Wilsson and Sundgren, 1998; Sinn *et al.*, 2010; Svartberg, 2002), predicting shelter dog behaviour after adoption (Valsecchi *et al.*, 2011; Mornement *et al.*, 2015), understanding the stability of behaviour across ontogeny and personality dimensions in puppies (Riemer *et al.*, 2014b; Riemer *et al.*, 2016; McGarrity *et al.*, 2015; Barnard *et al.*, 2016), and discovering the genetic basis of personality variation that can shed light on behavioural qualities important to tracing the domestication of dogs (Ilska *et al.*, 2017; Persson *et al.*, 2016). Through this burgeoning research, a large number of traits have been proposed and studied through a variety of different methods. Now, the field is in need of trying to find a common structure to the organisation of dog personality (Fratkin, 2017). Moreover, the predictive validity of personality assessments in dogs has been questioned, particularly in shelter dogs (Mornement *et al.*, 2015; Mohan-Gibbons *et al.*, 2012) and in some cases working dogs (Wilsson and Sundgren, 1998; Sinn *et al.*, 2010). Addressing these issues requires a closer look at how personality in dogs is studied, how personality traits are determined, and what advancements could be made.

5.1 Personality traits in dogs

Attempts at finding a common personality structure in dogs, such as the Five Factor model of human personality (McCrae and Costa, 1995), have not yet found consensus (Fratkin, 2017). Jones and Gosling (2005) summarised personality traits in dogs using seven dimensions: reactivity, fearfulness, sociability, responsiveness to training, aggression, dominance/submission and activity. Later, Fratkin *et al.* (2013) conducted a meta-analysis using the same framework, although decided to combine fearfulness and reactivity into a single fearfulness dimension. In puppies, McGarrity *et al.* (2015) found nine personality dimensions: activity, aggressiveness, boldness/self-assuredness, exploration, fearfulness/nervousness, reactivity, sociability, submissiveness, and trainability/responsiveness.

Other common categorisations of personality traits in dogs come from frequently used questionnaires and surveys, which require respondents to rate a dog's behaviour on a series of questions using ordinal rating scales. For instance, the Canine Behavioral Assessment and Research Questionnaire (C-BARQ; Hsu and Serpell, 2003) has been used in a variety of settings to learn about the behaviour of pet dogs (Asp *et al.*, 2015), shelter dogs (Duffy *et al.*, 2014; Barnard *et al.*, 2012), and working and service dogs (Serpell and Duffy, 2016; Foyer *et al.*, 2014). The CBARQ has evolved over the years, but now includes fourteen different subscales: stranger-directed aggression, owner-directed aggression, dog-directed aggression, dog rivalry, stranger-directed fear, nonsocial fear, dog-directed fear, touch sensitivity, separation-related behaviour, attachment of attention seeking, trainability, chasing, excitability, and energy. Other popular questionnaires are the Monash Canine Personality Questionnaire-Revised (Ley *et al.*, 2009), which evaluates dog behaviour with regard to five dimensions (extraversion, motivation, training focus, amicability and neuroticism), or the Dog Personality Questionnaire (Jones, 2008) that also uses five dimensions (fearfulness, aggression towards people, aggression towards animals, activity/excitability and responsiveness to training). Additionally, there have been questionnaires developed for more specific traits, such as the Dog Impulsivity Assessment Scale (Wright *et al.*, 2012) that investigates three facets of impulsivity (behavioural control, response to novelty and responsiveness), or the Highly Sensitive Dog questionnaire to investigate 'sensory processing sensitiy' (Braem *et al.*, 2017).

Personality and personality dimensions in dogs are also studied by means of direct behavioural observation, such as in test batteries, which are particularly common in animal shelters for determining the suitability of dogs to be rehomed. Mornement *et al.* (2014) developed the Behavioural Assessment for re-homing K9s (B.A.R.K) that consists of twelve subtests measuring five behavioural traits: anxiety, compliance, fear, friendliness and activity level. Similarly, Valsecchi *et al.* (2011) developed a temperament test for shelter dogs comprised of twenty-two subtests assessing sociability towards humans and conspecifics, playfulness, problem solving skills, trainability, possessiveness, and reactivity. Test batteries are, in addition, frequently used to evaluate the behaviour of potential working or service dogs. One of the most notable examples is the Dog Mentality Assessment developed by Svartberg and Forkman (2002), which is used by the Swedish Working Dog Association, and measures five dimensions (playfulness, curiosity/fearfulness, chase-proneness, sociability and aggressiveness).

5.2 Personality consistency in dogs

How consistent is personality in dogs? Fratkin *et al.* (2013) investigated the rank-order stability of behaviour through time across thirty-one different studies. Over an average inter-test time interval of 21 weeks, the average Pearson's correlation coefficient was $\rho = 0.43$. Fratkin *et al.* (2013) highlighted that this estimate of consistency or behavioural repeatability is similar to that in a meta-analysis across a wide range of taxa by Bell *et al.* (2009), who found an average ICC of 0.37. However, as noted by Nakagawa and Schielzeth (2010), Pearson's correlation is a measure of relative consistency rather than absolute consistency in behaviour. In fact, a Pearon's correlation coefficient of $\rho = 0.43$ indicates that only 19% (0.43^2) of behavioural variation at one time point in dogs can be explained by previous time points.

To see this, Figure 4 displays simulated data where the behaviour of one-hundred individuals has been measured across five occasions, with a correlation through time of $\rho = 0.4$. Figure 4 *(a)* displays the linear regression lines or *reaction norms* for each individual. While the slopes of many individuals are positive, there is considerable crossing of the regression lines across individuals. Figure 4 *(b)* displays raw data (black points) and reaction norms four randomly-selected individuals. Overall, while a correlation of $\rho = 0.4$ suggests weak to moderate consistency in behaviour through time, there are a number of other types of behavioural variation worth quantifying. Notably, behavioural plasticity (i.e. variation in the regression slopes) and individual differences in residual variance or 'predictability' (grey ribbons in Figure 4 *(b)*) may confer additional insights the behaviour of dogs. To my knowledge, only McGarrity *et al.* (2016) have assessed individual differences in average behaviour (i.e. personality) and behavioural change (i.e. plasticity) using hierarchical statistical models in military working dogs, although the authors found little evidence for significant behavioural variation in plasticity for the majority of behaviours studied.

More recent studies have quantified behavioural repeatability using the ICC. For example, Riemer *et al.* (2016) estimated the amount of absolute consistency in a number of personality traits at 6, 12 and 18 months of age in Border collies, finding an average ICC of 0.42, which is more comparable to estimates in other animals (Bell *et al.*, 2009). Moreover, Riemer *et al.* (2014a) found that measures of impulsivity using the DIAS scale, mentioned earlier, had high ICC values (mostly > 0.7) over a inter-test time interval of seven years. As the authors discuss, this may be because trait impulsivity is more supported by neurobiological findings than other personality traits. McGarrity *et al.* (2016) calculated the ICC for a number of behavioural traits in military work-

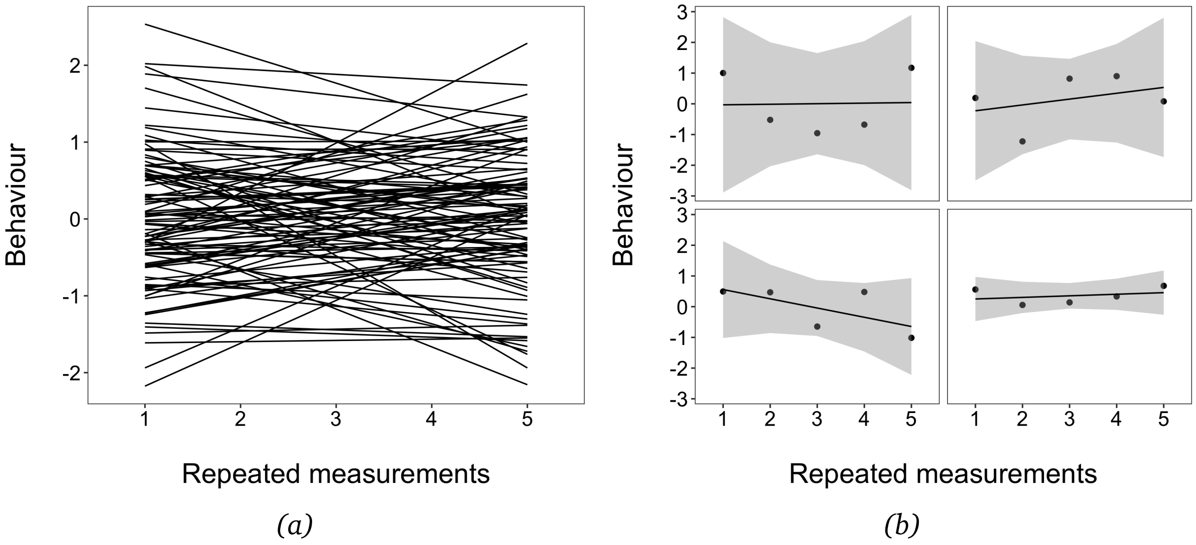


Figure 4: (a) Simulated reaction norms for one-hundred hypothetical individuals with a correlation of $\rho = 0.4$ across 5 repeated measurements, similar to that found in a meta-analysis Fratkin et al. *(2013). (b) Raw data (black points) and reaction norms for four randomly selected individuals. Shaded areas represent the residual variation around reaction norm estimates.*

ing dogs, using both behavioural rating (e.g. evaluating behaviour on Likert scales) and behavioural coding (e.g. measuring the frequency, duration or number of times a behaviour occurs) methods. Interestingly, the average ICC for the behavioural ratings was 0.31 whereas the average ICC for behavioural coding methods was only 0.15. As McGarrity *et al.* (2016) note, behavioural codings are more fine-grained than rating methods, and so may be more sensitive to behavioural variation through time.

Personality consistency has also been questioned because the predictive validity of personality assessments in a number of studies has been low. This is often the case when trying to predict the behaviour dogs across markedly different enviornments, such as the behaviour of shelter dogs after adoption (Mornement *et al.*, 2015; Mohan-Gibbons *et al.*, 2012; Poulsen *et al.*, 2010). Patronek and Bradley (2016) argue using simulation that up to half of assessments in shelters where dogs behave aggressively are likely to be false positives, because the base rate frequency of aggression outside of shelters is generally low (e.g. somehwere between 10 and 20% of dogs have shown aggression; Patronek and Bradley, 2016) and the sensitivity (proportion of correctly identified true positives) and specificity (proportion of correctly identified true negatives) of behavioural assessments in shelters are also expected to be low. Rayment *et al.* (2015) suggest that moving towards longitudinal and observational modes of assessment in shelters, rather than test batteries, may increase the ability to predict future dog behaviour. Personality has also been difficult to predict over ontogeny. In pet dogs, Riemer *et al.* (2014b) found little association between neonatal behaviour

(2-10 days old) and behaviour at 6-7 weeks of age or at 1.5-2 years old in Border collies. In military dogs, Wilsson and Sundgren (1998) report that puppy behaviour did not significantly predict adult performance on the same tests.

Nonetheless, certain behavioural assessments have been predictive of later behaviour. For working and service dogs, assessments that predict a binary pass or fail result on a test from earlier behaviour have had greater predictive accuracy. Sinn *et al.* (2010) found some predictive accuracy in a military working dog test, after combining test results into aggregated behavioural variables. Harvey *et al.* (2016b), furthermore, developed a behavioural test battery for potential guide dogs, and found a number of behaviours (e.g. responding quicky to a "down" command) and composite, principal component scores (e.g. low values for distraction or fear/anxiety components) to be significantly predictive of qualification as a guide dog (see also Harvey *et al.*, 2017).

5.3 Approaches to animal personality: where do dogs sit?

The vast majority of studies in dogs personality have followed a latent variable approach, as explicated in section 4.2.2. Multivariate data is relatively easy to collect on dogs due to their accessibility, whether using questionnaires or standardised behavioural tests. For instance, the C-BARQ is composed of one-hundred different items pertaining to the fourteen subscales or factors mentioned earlier. Thus, latent variable models that can reduce multivariate datasets into a set of smaller variables that explain a large proportion of the variance are essential. However, formative models, in particular principal components analysis, are considerably more popular than reflective models, and confirmatory models. I conducted a short *Web of Science* database search for articles published between January 2016 and August 2017 using the terms 'dog' and 'personality', and recorded the topic of the study, the data collection method used and the statistical methods applied to identify personality traits of interest. Twenty-seven studies were found, and Table 1 summarises the thirteen studies that aimed to determine personality dimensions underlying behaviour or confirm previous findings (studies that did not attempt to determine or replicate previous dimensions were removed).

Nine of the thirteen studies in Table 1 used principal components analysis to derive personality dimensions from the behavioural data. In some cases, these studies had *a priori* hypotheses that could have been tested using confirmatory approaches. For example, Harvey *et al.* (2017) developed a questionnaire to assess potential guide dog behaviour that targeted seven personality traits. Although the principal components analysis and other exploratory methods found seven components, a confirmatory factor analysis would have been a more powerful approach for ascertaining the validity of the questionnaire in assessing the targeted personality traits. As discussed previously, principal components analysis will always return components that explain the greatest amount of variation in the data (Budaev, 2010; Beaujean, 2014) and, thus, the null hypothesis that no underlying, lower-order variables explain the data cannot be adequately tested. In one study, exploratory factor analysis was used (Nagasawa *et al.*, 2016). The only study to use a confirmatory factor analysis during this period was Barnard *et al.* (2016), who attempted to replicate in puppies the four-factor personality structure found in adults dogs by Ley *et al.* (2008). Barnard *et al.* (2016) instead demonstrated that a four-factor structure did not fit the data as well as a five-factor structure, using measures of model fit such as the root mean squared error of approximation and the comparative fit index.

Apart from the studies in Table 1, some authors have developed personality assessments using a mixture of exploratory and, subsequently, confirmatory methods. Jones (2008) developed the Dog Personality Questionnaire through a process of applying exploratory and confirmatory factor analysis. Moreover, Ley *et al.* (2008) developed the Monash Canine Personality Questionnaire using principal components analysis, but later revised the questionnaire (Ley *et al.*, 2009) after structural equation modelling suggested that the previous structure could not be replicated. Such revisions through applying confirmatory models could be fruitfully applied to other instruments measuring personality in dogs, or in meta-analyses.

Table 1: All publications between January 2016 and August 2017 assessing personality traits in dogs from a Web of Science search. Abbreviations used: PCA (principal components analysis); EFA (exploratory factor analysis); CFA (confirmatory factor analysis).

Reference	**Topic**	**Data collection**	**Statistical methods**
Harvey *et al.* (2017)	Predicting guide dog qualification from 5, 8 and 12 month behaviour	Questionnaire	PCA
Diverio *et al.* (2017)	Association between avalanche search dog-handler behaviour and performance on a simulated trial	Focal animal sampling while working	PCA
Braem *et al.* (2017)	Developing the 'Highly Sensitive Dog' questionnaire to investigate sensory processing sensitivity	Questionnaire	PCA (for sensory processing sensitivty questions only)
Barnard *et al.* (2017)	Personality in 2 month old dogs in an open field test	Standardised behavioural assessments	Hierarchical cluster analysis after EFA assumptions not met
Szánthó *et al.* (2017)	Developing the Dog Emotional Reactivity Survey to investigate empathy in dogs	Questionnaire	*A priori* subscale construction and checks of internal consistency
Sundman *et al.* (2016)	Comparing behavioural traits in pet/conformation and working retrievers	Standardised behavioural assessments	PCA
Harvey *et al.* (2016a)	Investigating rearing environment and behaviour at 5, 8 and 12 months old in potential guide dogs	Questionnaire	PCA
McGarrity *et al.* (2016)	Predicting working dog performance from behavioural rating and coding methods	Standardised behavioural assessments	PCA
Hoummady *et al.* (2016)	Comparing human and dog personality, and performance in working tasks	Standardised behavioural assessments	PCA
Barnard *et al.* (2016)	Comparing subjective rating and behavioural coding methods in an open field test with 2 month old dogs	Standardised behavioural assessments & questionnaire	Hierarchical cluster analysis and CFA
Fadel *et al.* (2016)	Investigating trait impulsivity across breeds and working/show lines	Questionnaire	PCA (to replicate previous DIAS components)
Nagasawa *et al.* (2016)	Comparing behavioural traits of dogs in the United States and Japan	Questionnaire	EFA
Harvey *et al.* (2016b)	Predicting guide dog qualification from 5 and 8 month behaviour	Standardised behavioural assessments	PCA

In addition, only one study has assessed the assumptions of latent variable models as mentioned in section 4.3.1. van den Berg *et al.* (2010) assessed measurement invariance using an item response theory model (a confirmatory, reflective latent variable model for ordered categorical manifest variables). The authors assessesed whether the stranger-directed aggression subscale/factor of the C-BARQ was measurement invariant (i.e. had the same structure) in German shepherds, Labrador retrievers and golden retrievers, and in different sex and neuter status groups within breeds. Although some violation of measurement invariance was found, the authors argued that it was small and inconsequential. Ideally, confirmatory modelling should strive to include tests of measurement invariance and other assumptions, such as local independence (section 4.3.1), when possible to ensure that explaining dog behaviour as a function of certain personality traits is warranted.

The operational approach has rarely been applied in studies of dog personality, although McGarrity *et al.* (2016) used hierarchical statistical models to assess both personality and behavioural plasticity in a number of behavioural traits. Individual differences in the residual variance, or behavioural predictability, have never been evaluated in dogs, to my knowlegde. Nonetheless, this topic is central to testing whether dogs vary in their intra-individual behavioural consistency, as hypothesised (Fratkin *et al.*, 2013). Operational approaches would be particularly useful in settings where longitudinal modelling is necessary, such as how dogs behave through time over ontogeny or at shelters.

Finally, network analysis has never been applied to understand dog behaviour or behavioural phenotypes in animals, generally. While network analysis is, currently, largely an exploratory method (Epskamp *et al.*, 2016b), the emphasis on understanding causal connections between behavioural, cognitive and affective components (inferred from conditional independence relationships) allows one to generate more specific hypotheses about the organisation of behaviour. Given the diverse number of personality dimensions that have been reported in dogs, network analysis may offer new insights into the causal relationships that exist between different behavioural variables in personality traits that show replicability, and how those causal relationships develop through time or ontogeny.

6 Aims of the thesis

Broadly, the aims of this thesis were to:

- Evaluate the conceptual and methodological issues involved in making inferences about personality and personality traits in dogs.
- Advance understanding of dog personality through time and across contexts.
- Propose new directions for the study of personality in dogs.

Paper I took a latent variable approach to studying personality traits in dogs, and evaluated whether the assumptions of local independence and measurement invariance in confirmatory, reflective latent variable models were satisfied using data on aggressiveness towards people and dogs in a population of shelter dogs. Measurement invariance was assessed in different sex and age groups. **Paper II** applied an operational approach to study personality, plasticity and predictability in shelter dogs' reactions to meeting unknown people at a shelter. Lastly, **Paper III** demonstrated how network analysis can be used to understand the organisation of behavioural phenotypes in police dogs, and how a network perspective encompasses, and can clarify our understanding of, animal personality.

7 Materials & Methods

Papers I and **II** used behavioural assessment data from Battersea Dogs and Cats Home, an animal shelter in the United Kingdom that cares for thousands of dogs per year. **Paper III** analysed data on police patrol and detection dogs in Norway. Details about the dogs and the data collection methods for the papers are summarised separately below.

7.1 Shelter dogs

Data from all dogs (N=4,990) being cared for by Battersea Dogs and Cats Home during 2014 (including those arriving to the shelter before, or departing after, 2014) were extracted with the shelter's permission from the computer database. **Paper I** used data on a sample of N=4,743 dogs and **Paper II** used data on a sample of N=3,263 dogs (full demographic details are reported in the papers). In both papers, all dogs were at least 4 months old because younger dogs were often housed in different kennels to older dogs and may have been limited in their interactions if still unvaccinated. While dogs were of a variety of breeds, breed differences in behaviour were not studied because the identification of breeds in shelter dogs is unreliable (Olson *et al.*, 2015; Voith *et al.*, 2013).

The shelter has three rehoming centres: a high-throughput, urban centre based at Battersea, London with capacity for approximately 150-200 dogs; a semi-rural/rural centre based at Old Windsor with capacity for approximately 100-150 dogs; and a rural centre based at Brands Hatch with capacity for approximately 50 dogs. Each dog's behavioural assessment is recorded in a custom computer system (see below for details). The kennels varied within and between the different rehoming centres, but were usually 4m x 2m, with a shelf and bedding alcove (see also Owczarczak-Garstecka and Burman, 2016). Dogs were generally housed individually for safety reasons, unless two dogs arrived into the shelter from the same home and it benefited them to share a kennel. All dogs had access to runs at the back of the kennel for at least part of the day. Dogs received a variety of social and sensory stimulation throughout the day, including daily socialising or training sessions with staff and volunteers, toys, music played in the kennel block areas, and access to quiet 'chill-out' rooms.

7.1.1 Observational behavioural assessment

The shelter uses an observational and longitudinal behavioural assessment. The core part of the assessment evaluates dog behaviour in 9 contexts: *Handling*, *Kennelling*, *Interactions with familiar people*, *Interactions with unfamiliar people*, *Out of kennel*, *Eating food*, *Interactions with toys*, *Interactions with female dogs*, *Interactions with male dogs*. For each context, trained shelter employees record behavioural observations using qualitative behavioural ethograms specific to that context in a custom computer system. Observations are carried out on a near-daily basis, or as frequently as the dog is observed in a particular context. The ethograms have between 10 and 16 behavioual codes (depending on the context). The codes are mainly adjectives with associated behavioural descriptions/definitions, and shelter staff choose one behavioural code that best describes the dog's behaviour in the particular context on that occasion. The ethograms are arranged as a scale into green, amber and red codes that reflect a dog's suitability for adoption: green behaviours pose no problems for adoption, amber behaviours suggest dogs may require some training to facilitate successful adoption but do not pose a danger to people or other dogs, and red behaviours suggest dogs needed training or behavioural modification to facilitate successful adoption and could pose a risk to people or other dogs. Multiple shelter staff could fill out observations for each dog.

In **Paper I**, we analysed aggressive behaviour towards people and dogs across the different shelter contexts. In each context, the red-category behavioural code *Reacts to people/dogs aggressive* was the most severe, defined as 'Growls, snarls, shows teeth and/or snaps when seeing/meeting other people/dogs, potentially pulling or lunging towards them'. Reactive and aggressive behaviour is distinguished from *Reacts to people/dogs non-aggressive,* defined as 'Barks, whines, howls and/or play growls when seeing/meeting other people/dogs, potentially pulling or lunging towards them'. The *Kennelling* and *Out of kennel* contexts were each split into two contexts, since aggressive behaviour could be recorded to either people and dogs in those contexts (full details/descriptions of the finall 11 contexts are reported in **Paper I, *Table 2***). We aggregated the data for each dog into a binomial variable, where 1 = the dog had a *Reacts to people/dogs aggressive* observation recorded at least once in a particular context during their time at the shelter.

For **Paper II**, we applied an operational approach to behaviour only in the *Interactions with unfamiliar people* context, which had an ethogram of 13 different behavioural codes, ranging from *Friendly* (i.e. 'the dog initiates interactions with people in an appropriate social manner) to *Reacts to people aggressive'*, as defined above (the full ethogram is reported in **Paper II, *Table 2***). Due to most behavioural observations being green codes, the scale was reduced into a 6-category ordinal scale representing 4 green codes, and 2 codes aggregating the amber and red codes, respectively. Thus, higher scale codes reflected less sociable responses. The analyses were focused on behavioural change during the first month of arrival to the shelter (arrival day 0 to day 30), since the average number of days spent at the shelter is ususally around 25-30 days and observations were much less frequent after day 30.

7.2 Police dogs

Paper III collected questionnaire data on N=171 police dogs in Norway, in conjunction with the Norwegian Police University College in Kongsvinger. The responses analysed included 117 patrol dogs (91 German shepherd dogs; 22 Belgian malinois; 1 rottweiler; 1 giant schnauzer; 1 Belgian tervueren; 1 unrecorded breed) and 54 detection dogs (17 labradors; 12 flat coated retrievers; 8 German shepherd dogs; 8 springer spaniels; 2 Belgian malinois; 2 Welsh springer spaniels; 1 German shepherd dog x Belgian shepherd dog; 1 labrador x German pointer; 1 cocker spaniel; 1 Nova Scotia duck-tolling retriever; 1 unrecorded breed). The dogs mostly uncastrated (n = 117) and male (n = 149). The responses were completed by 117 police dog handlers (79 male; 17 female) between 28 and 57 years old, and with between 1 and 30 years of experience.

7.2.1 Questionnaire

The questionnaire was developed with senior members of the Norwegian Police University College to investigate the working and non-working lives of the police dogs and their handlers. For **Paper III**, the personality section of the survey was analysed, which asked respondents to rate their dogs on 43 adjective-based and situational descriptors targeting specific motivational and behavioural characteristics. The responses were recorded on 5 point rating scales, 1 = 'Strongly disagree' to 5 = 'Strongly agree', where 3 = 'Neutral'. Participants could also choose 0 = 'Not relevant/I do not know'. Due to the low sample size, we carefully screened the data to avoid to retain only those descriptors that were most reliably recorded. This included removing descriptors that: received more than 10% of missing responses, had little variation, were highly correlated (i.e. > $\rho = 0.8$) indicating redundancy (i.e. retaining only one descriptor rather than multiple, highly correlated descriptors), or descriptors where pseudo-replication was a potential problem for those handlers who filled out multiple questionnaires on different dogs (*N*=44). The final analyses were conducted on 20 descriptors.

7.3 Validity & inter-rater reliability

For the shelter dog data, the reliability and validity of the observations could not be ascertained directly while maintaining the anonymity of the employees recording the observations in the computer system. Instead, separate video-coding sessions were run with *N*=93 staff members across the three different rehoming centres who were trained in completing behavioural observations on the dogs. Experienced canine behaviourists at the shelter recorded videos of 14 randomly-chosen behaviours (approximately 30 seconds each), 2 from each of 7 assessment contexts (the *Interactions with familiar* and *unfamiliar people*, and *Interactions with female* and *male dogs*, contexts were combined into single *Interactions with people* and *Interactions with dogs* contexts, respectively). Shelter employees watched the videos in small groups (usually between 5 and 10 people in each session), and recorded on answer sheets after viewing each video which ethogram code best described the behaviour. Employees answered individually, and were only allowed to watch the videos once. For **Paper I**, we analysed the responses to two videos: one illustrating *Reacts to people aggressive* in the *Interactions with people* context, and another illustrating *Reacts to dogs aggressive* in the *In kennel towards dogs* context. For **Paper II**, the two videos in the *Interactions with people* context illustrated a *Reacts to people aggressive* response and a *Reacts to people non-aggressive*.

No specific checks of validity or reliability were made in **Paper III**. It was unlikely to find other police dog handlers who knew the dogs well enough in order to assess inter-rater reliability, and the questionnaires were not issued to the handlers again to assess intra-rater reliability due to time constraints. While some measures of validity were upheld, such as convergent validity (i.e. whether descriptors expected to correlate with each other are in fact correlated) or divergent validity (i.e. whether descriptors not expected to correlate with each other do not in fact correlate), these validity metrics are not clearly relevant in a network perspective. Common traditional notions of validity are tightly linked to a reflective latent variable view of scientific constructs (Cramer, 2012; Borsboom, 2005; Boag, 2015). In other words, variables correlate with each other in particular ways because they reflect the same, or different, underlying scientific constructs.

7.4 Data analysis

All data analysis was conducted in the R statistical environment (R Core Team, 2017).

7.4.1 Validity & inter-rater reliability

For **Paper I** and **Paper II**, validity was assessed by the percentage of shelter employees who selected the correct behavioural code (as determined by the experienced behaviourists filming the videos) to describe the behaviours in the videos. Inter-rater reliability was assessed using the consensus statistic in the *agrmt* package (Ruedin, 2016), which is based on Shannon entropy and measures the amount of agreement in ratings of ordered categorical data.

7.4.2 Missing data

Missing data was handled using multiple imputation, rather than listwise deletion or mean substitution (Rubin, 1976), using the *Amelia* package (Honaker *et al.*, 2015). In **Paper I**, missing data occurred when dogs received no observations in a particular context throughout their stay at the shelter. For most contexts, the missing data rate was between 0.06% and 5%, although the *Interactions with female dogs* and *Interactions with male dogs* categories had 17% and 18% missing values, respectively, because structured interactions with other dogs did not arise as frequently). In **Paper III**, we imputed missing data for descriptors that had no more than 5% of missing responses, which occurred for. For **Paper II**, we did not impute missing values for days in which dogs had no observations, since it was difficult to determine whether the dog had met an unfamiliar person on that day, and the observation had not been recorded, or if the dog had just not met any unfamiliar people that day. Since all the dogs in the sample of data analysed had some behavioural observations, we chose not to use multiple imputation.

7.4.3 Inferential models

Paper I: The 11 aggression contexts were first analysed with structural equation modelling using the *lavaan* package (Rosseel, 2012), and the results were combined across the imputed data sets using functions in the *semTools* package (semTools Contributors, 2017). Two latent variables were specified: one underlying the seven contexts where *Reacts to people aggressive* codes were recorded (*Handling, In kennel towards people, Out of kennel towards people, Interactions with familiar people, Interactions with unfamiliar people, Eating food, Interactions with toys*), and the other to the four contexts where *Reacts to dogs aggressive* codes were recorded (*In kennel towards dogs, Out of kennel towards dogs, Interactions with female dogs, Interactions with male dogs*). We compared a model with orthogonal latent variables, to one where the latent variables were allowed to correlate. Model fit was ascertained using the comparative fit index (CFI) and Tucker Lewis index (TLI), where values > 0.95 indicated excellent fit, as well as the root mean squared error of approximation (RMSEA) where values < 0.06 indicated good fit. The assumption of local independence (i.e. manifest variables are uncorrelated after conditioning on the latent variable) was tested by specfying pre-defined covariances between aggression contexts that were believed to have close temporal and spatial relationships. For example, the *Handling* context could be closely preceded by a number of contexts (e.g. *Interactions with familiar people*), which may be revealed as a violation of local independence.

The assumption of measurement invariance across different sex and age groups was tested by using hierarchical Bayesian logistic regression models for each trait separately (i.e. aggressiveness towards people and dogs contexts, respectively) written in the probabilistic programming language Stan (Carpenter *et al.*, 2016) and run in R through the *rstan* package (Stan Development Team, 2016). The hierarchical logistic regression models modelled the probability of aggression in different contexts as a function of individual latent aggression levels (i.e. 'random intercepts'), the different aggression contexts, sex, and age groups (4 to 10 months; 10 months to 3 years; 3 to 6 years; over 6 years). Measurement invariance was violated if there were significant interactions between sex or age groups and the aggression contexts. In addition, the models took into account a number of other predictors that were not inferentially interpreted: 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). Behavioural repeatability was assessed by calculating the ICC. Models including interaction terms were compared to simpler models without interactions using the widely applicable information criterion (WAIC; Watanabe, 2010), which indicates the out-of-sample predictive accuracy of statistical models.

Paper II: the framework of behavioural reaction norms was applied using a hierarchical Bayesian ordinal probit model written in the Stan language and run in R through the *rstan* package (e.g. for similar models, see Liddell and Kruschke, 2015; Foulley and Jaffrézic, 2010). Each dog's behaviour was modelled as a function of their average behaviour across their observations (i.e. personality), a linear and quadratic function of day since arrival to the shelter (i.e. linear and quadratic plasticity), and the residual variance around the reaction norm estimates (i.e. predictability). Correlations between these dog-specific parameters were estimated. Behavioural repeatability was calculated using the ICC. Because the between-individual differences was a function of day since arrival, the ICC pertained to particular days since arrival. The ICC was evaluated on days 0 (arrival day), 8 and 15. Furthermore, we reported the 'cross-environmental correlation' defined by Brommer (2013), which allows one to assess the rank-order stability of individuals' reaction norm estimates between specific time points. The cross-environmental correlation was evaluated also at days 0, 8 and 15. The models took into account the same predictor variables as **Paper I**, reported above. Model selection was also performed using the WAIC, by comparing the full model estimating all personality, plasticity and predictability parameters to a series of simpler models.

Paper III: Networks of conditional independence relationships (or partial correlations/Gaussian graphical models) between the motivational and behavioural descriptors of police dog behaviour were constructed using the *qgraph* package (Epskamp *et al.*, 2012). The networks were estimated for patrol and detection dogs separately, using the polychoric correlation matrices. The networks employed L_1 lasso penalties (i.e. least absolute shrinkage and selection operator), where the matrix of partial correlations were regularised so that partial correlations near zero were shrunken towards zero. The amount of regularisation was determined by a parameter λ, which was selected by the model that minimised the extended Bayesian information criterion (EBIC), implemented in the *qgraph* package. The networks were analysed by computing the betweenness and strength centrality statistics for each node, which indicated descriptors that were important for maintaining network connectivity. Nodes or descriptors with high betweenness values acted as mediators between indirectly connected nodes, and nodes with high strength values had stronger correlations with other descriptors. To assess how sensitive the networks were to sample size or the number of descriptors included in the network, stability analyses were conducted using non-parametric bootstrapping (in the *bootnet* package; Epskamp *et al.*, 2016a) that constructed networks using different numbers of dogs and descriptors, and compared the results to the original networks (Epskamp *et al.*, 2017a). Moreover, non-parametric bootstrapping was also used to compare the patrol and detection dogs networks, and the differences in the descriptors' centrality values between the two networks were compared using Cliff's delta, a measure of effect size (Torchiano, 2016).

7.5 Ethical approval

Ethical approval from the Regional Ethics Committee in Norway was not required for the papers in this thesis because the studies did not involve handling or experimenting on animals. A written agreement was signed with the shelter permitting use of the data for the analyses and publication. Approval from the Norwegian Social Science Data Services was acquired for **Paper III** for the processing of personal data (approval no. 44121).

7.6 Data accessibility

The raw data, R code, Stan model code and supplementary materials for **Paper I** can be found at `https://github.com/ConorGoold/GooldNewberry_aggression_shelter_dogs`, and for **Paper II** at `https://github.com/ConorGoold/GooldNewberry_modelling_shelter_dog_behaviour`. The data, R code, and supplementary materials for **Paper III** can be found alongside the online publication.

8 Results and Discussion

8.1 Assumptions of latent variable approaches

The results of **Paper I** demonstrated that the structural equation model representing inter-context aggressive behaviour towards people and dogs with two correlated latent variables fit the data well (CFI: 0.96; TLI: 0.95; RMSEA: 0.03). This supports previous research in dogs separating aggressiveness into people-directed and dog-directed traits (e.g. Hsu and Serpell, 2003; Jones, 2008). Moreover, aggression across contexts was moderately repeatable for both traits, although repeatability was higher for the aggression towards people (ICC = 0.48) observations than aggressiveness towards dogs (ICC = 0.30).

Nonetheless, violations of local independence and measurement invariance were also found. The structural equation model including covariances between certain aggression contexts improved the overall model fit (CFI = 0.98; TLI = 0.97; RMSEA: 0.03). For example, significant negative relationships were observed between aggression in the *Handling* context, and the *In kennel towards people* and *Interactions with unfamiliar people*, respectively. This likely occurred because if dogs showed aggression towards people in the latter two contexts, shelter staff would be less likely to handle the dog or handle the dog more carefully and, therefore, the dogs would actually be *less* likely to show aggression in the *Handling* context. This highlights the problems with averaging, or taking a sum, over items assumed to the measure the same personality trait, as if often conducted in questionnaires on dogs (e.g. C-BARQ subscales are summarised by the average of the items; Asp *et al.*, 2015). If responses to those items violate the assumption of local independence, there may not be clear relationship between the level of an underlying trait and the average of items measuring that trait. Violations of local independence are, further, of particular concern to test batteries, where the behaviour of dog and human tester may be influenced by preceding test responses. In shelter dogs, the stress of a dog being taken from their kennel by unfamiliar people undergoing a test battery could obscure the accurate assessment of a personality trait if test responses are influenced by the stress of the dog, as well as the targeted trait being measured (Rayment *et al.*, 2015).

For the assumption of measurement invariance, models that took into account interactions between the aggression contexts and age/sex groups had greater out-of-sample predictive accuracy (lower WAIC values) than simpler models without interaction parameters. Violations of measurement invariance were observed for both aggressiveness towards people and dog contexts. For instance, female dogs had greater odds than males of showing aggression in the *Out of kennel towards people* and *Interactions with unfamiliar people* contexts compared to other contexts. Female dogs also had similar odds of aggression towards female and male dogs, whereas males were significantly more likely to show aggression towards other males than females. Between age groups, dogs up to 6 years old had greater odds of showing aggression in the *In kennel towards people* and *Interactions with unfamiliar people* contexts, whereas dogs over 6 years old were most likely to show aggression in the *Handling* context and had increased odds of showing aggression in both *Eating food* and *Interactions with toys* contexts relative to younger dogs. Demographic differences in aggression in dogs are of particular interest to researchers, governing bodies and the lay-public alike (Casey *et al.*, 2014; Orritt, 2015), so evaluating whether those differences can be reliably interpreted as differences in an aggressiveness trait is paramount to avoid misleading conclusions.

Violations of these assumptions indicate theoretical problems for the explanation of behavioural responses as a function of underlying latent traits. In psychology, evaluating local independence and measurement invariance are particularly important for fair assessment in education or psychopathology. Wicherts and Dolan (2010), for instance, illustrate violations of measurement invariance in intelligence tests between majority and minority ethnic children in the Netherlands, finding biases against the minority children leading to the achievement of lower IQ scores. Yet, interpreting the lower IQ scores as a function of IQ is obfuscated by a lack of measurement invariance across ethnicities. Local independence is often violated in cases where redundancy exists in a particular set of items measuring a construct. For example, Edelen and Reeve (2007) found certain violations of local independence in a nineteen-item depression scale, specifically between positively-worded items. While local independence and measurement invariance may be too stringent in many cases (Markus and Borsboom, 2013), or their practical implications may be difficult to determine, their violation does not have to invalidate an assessment, but can be used to refine (e.g. through the removal of locally dependent items) or amend inferences on test scores (e.g. if particular patterns of measurement variance are known).

In animal personality more broadly, local independence and measurement invariance have not been investigated directly, beyond the assessment of measurement invariance by van den Berg *et al.* (2010) in dogs. However, the application of confirmatory, reflective latent variable models is becoming increasingly more popular in areas such as behavioural and evolutionary ecology (Araya-Ajoy and Dingemanse, 2014; Martin and Suarez, 2017; Dochtermann and Jenkins, 2007), and evaluating their assumptions could advance our understanding of which personality traits and dimensions can be reasonably compared across studies, and across species. For instance, an interesting application of testing measurement invariance would be the assessment of species-specific differences in personality traits (Koski, 2011; Uher, 2011) to determine whether the measurement relationships between the latent and manifest variables are the same in different species and, therefore, species can be reasonably compared on the same trait. An influencing factor for both violations of local independence and measurement invariance is not conditioning, statistically, on other latent variables that may influence the manifest variables. In **Paper I**, fearfulness could influence the expression of aggressive behaviour in certain contexts (e.g. when meeting unfamiliar people), and different sex or age groups may have differential levels of fearfulness. Since personality traits themselves tend to correlate with each other (i.e. ‘behavioural syndromes’; Sih *et al.*, 2004), it may be difficult to avoid the combined effects of different latent variables on the same manifest variables.

8.2 Beyond personality

Difficulties with applying a latent variable approach may be partly alleviated by taking an operational approach to studying personality, where greater emphasis is placed on assessing between- and within-individual differences through time in single behavioural measures. The results of **Paper II** demonstrate that taking into account personality, plasticity and predictability simultaneously in dog behaviour improved the out-of-sample predictive accuracy of the statistical models considerably. Indeed, modelling individual differences in behavioural predictability appeared particularly important from improving model performance, indicating that individual differences in residual variances could be an important avenue of future research in dogs. Indeed, predictability represents what many authors view as behavioural consistency (Fratkin *et al.*, 2013), since it represents within-individual variation through time. Nonetheless, behavioural consistency is frequently inferred by correlation coefficients across different time points, which instead represent the stability of between-individual differences.

Dogs’ responses to meeting unfamiliar people largely fell into the *Friendly* category (63.5% of responses), although sociability towards unfamiliar people increased across days since arrival. Moreover, the quadratic effect indicated that behavioural change, or plasticity, tended to be greatest early after arrival rather than later, across all dogs. Some studies on shelter dogs have reported an improvement in stress-related behaviours through time (Stephen and Ledger, 2005) whereas others have indicated an increase in the probability of problem behaviours, such as aggression towards unfamiliar people (Kis *et al.*, 2014). The results in **Paper II** are the first to systematically model individual differences in plasticity in shelter dogs, however, whereas previous research has examined group-level behavioural trajectories only. By doing so, variation in within-individual behavioural variation may be obscured. For example, while most dogs' responses to meeting unfamiliar people became more sociable with time at the shelter, that was not the case for all dogs: some dogs showed little behavioural change, and others showed a decrease in sociability over time.

Behavioural repeatability, or the amount of variation explained by between-individual differences, increased after the first week in the shelter. The ICC on arrival day was 0.22, whereas eight days after arrival it increased to 0.33 (but changed little between day 8 and 15). This implies that differences between dogs were more stable after the first week of arriving to the shelter, and therefore shelters may benefit by waiting a week before making any clear decisions about a dog's typical behaviour. Moreover, the cross-environmental correlation indicated that the rank-order stability between arrival and day 15 in the shelter was only moderate, implying that the most sociable dogs on arrival were not the most sociable after a couple of weeks in the shelter.

An important finding in **Paper II** was that there was substantial uncertainty in the individual-level reaction norm estimates. While there was a relatively large sample size of dogs (> 3000) to inform the group-level results, dogs had on average only 5.9 (standard deviation = 3.7; range = 1 to 22) recorded observations of meeting unfamiliar people. Consequently, estimation of each dog's personality, plasticity and predictability parameters entailed substantial uncertainty. On one hand, more observations per dog would enable more accurate predictions, although the uncertainty around the reaction norm estimates is also a function of the amount of between-individual variation. On the other hand, this number of observations is typical of the amount of data shelters can reasonably be expected to collect on dogs in their care. Thus, uncertainty in predicting any one dog's behaviour is likely to be the norm. A Bayesian approach is ideal for these circumstances because it quantifies uncertainty in parameter estimates given the data (i.e. provides a posterior probability distribution), and has a number of advantages over the interpretation and construction of frequentist confidence intervals, which are not posterior distributions (Kruschke, 2014).

Should the term personality be reserved for differences in average behaviour only, or could it act as an overarching term for different types of behavioural variation? Personality is typically considered differences in how animals behave on average. In dogs, questionnaire data aims to obtain responses about the most probable behavioural responses of dogs, which may be inferred through Likert scale responses on the frequency of behaviour (e.g. ‘Never’ to ‘Always’). At the same time, some authors use personality to encompass both between- and within-individual variation. For instance, Fratkin *et al.* (2013) suggested that the temporal consistency of behaviour may be a personality trait itself, with some dogs being more consistent than others. However, this notion of temporal consistency indicates a within- rather than a between-individual phenomenom, whereas estimates of behavioural consistency in animal personality, and the meta-analysis of Fratkin *et al.* (2013), typically concern the stability of between-individual differences. As discussed in the Introduction, behavioural repeatability can be measured a number of different ways, and researchers have emphasised the correct interpretation of repeatability estimates (e.g. Biro and Stamps, 2015).

8.3 Applying the network perspective

The results of the network analyses in **Paper III** indicated a number of strong conditional dependence relationships between functionally-related behaviours. For example, motivational and behavioural descriptors related to aggression (‘Dog aggressive’, ‘Strong tendency to growl at strangers’ and ‘Food aggressive’) correlated positively, as did descriptors related to sociability or trainability (‘Socially attached to you’, ‘Willing to please’, ‘Recalls’, ‘Willing to give you a toy’). The networks also demonstrated strong correlations between descriptors similar to a shyness-boldness dimension in working dogs. In particular, the descriptors ‘Playful’, ‘Curious’, ‘Fearless’ and ‘Socially attached to you’ are somewhat similar to the correlated factors found by Svartberg and Forkman (2002), comprising playfulness, curiosity/fearlessness and sociability dimensions.

The difference here from the latent variable approach above is that there is no assumption of an underlying variable causing these correlations. Rather, the conditional dependencies between the descriptors are taken as signs of causal relationships between the different behavioural and motivational characteristics, which can generate new hypotheses about the organisation of behaviour. For instance, dogs with high agreement scores for ‘Willing to give you a toy’ may be more likely to have high scores for ‘Recalls’ for a wide range of reasons, such as handlers that can train a dog to do one behaviour also train their dog to do the other behaviour, or that dogs are trained to recall (i.e. coming back when called) using a toy as a reward that subsequently entails the dog relinquishing a toy to their handler. Such simple relationships may be the cause of widespread correlations in behavioural data on dogs and other animals, rather than the presence of underlying but unobserved personality dimensions. It is also worth noting that the networks indicated that some variables that shared normal pairwise correlations were not correlated in the networks of conditional independence relationships. 'Curious' and 'Socially attached to you' were positively correlated (0.41 and 0.40 for patrol and detection dogs, respectively), but not directly related in either network. This suggests that their pairwise correlation was due to common mediating variables, rather than any direct relationship between being curious and socially attached to the handler.

A key aspect of **Paper III** was in determining the centrality of different descriptors in the network. Centrality statistics come in a variety of forms, but all indicate how important nodes in the network are for determining network structure. In both patrol and detection dogs, the 'Playful' descriptor had particularly high betweenness and strength centrality values. Considering that playing is used as a reward in the training of police dogs in Norway, it is perhaps unsurprising that variation in a dog's rating on the 'Playful' descriptor could have widespread influences on the organisation of behaviour more generally. Patrol and detection dogs differed in centrality values for certain descriptors. For example, the 'Curious' descriptor was more central in the patrol dog network, as were 'Good at walking on slippery surfaces' and 'Active and nimble'. By contrast, in the detection network, more task-specific descriptors were more central than in the patrol dog network, such as 'Able to stay focused during searches' and 'Gives up searches quickly'.

8.4 Limitations & future directions

One general limitation common to all the approaches is the need for data on a large number of subjects. For latent variable models, researchers recommend at least a 1:5 variable-to-subject ratio (Beaujean, 2014), although this will be dependent on the specific study and complexity of the model. Similarly, hierarchical statistical models used by the operational approach may require a large number of subjects to estimate the between-individual differences accurately, particulary individual differences in behavioural plasticity. For instance, in one scenario, Martin *et al.* (2011) recommend at least 200 total observatins in the data set as a useful rule of thumb (e.g. 20 observations on 10 individuals; see also Dingemanse and Dochtermann, 2013). Nonetheless, this amount of data may be easier to collect on dogs than in other species.

Although in its infancy, the network approach is also sensitive to small sample sizes, although there are few rules of thumbs for an optimal variable-to-subject ratio (Epskamp *et al.*, 2017b). Rather, work has been invested in estimating networks that are conservative, by using regularisation (i.e. shrinking weak correlations towards zero; Epskamp *et al.*, 2017a), so that the removal of spurious connections in the network are prioritised. We demonstrated this in **Paper III**, where the sample sizes for the police patrol and detection dogs were relatively small (117 and 54, respectively). In addition, we analysed the stability of the network results by using non-parametric bootstrapping (Epskamp *et al.*, 2017a). The results indicated that the patrol dog network results were relatively resilient to changes in sample size and the number of nodes in the network, although the detection dog network results were more sensitive in networks of differing sample sizes. Given issues with reproducibility in science more generally, getting the network perspective off 'on the right food' is essential.

A more specific limitation to **Paper I** and **Paper II** were estimates of validity. For both papers, validity was evaluated as how well the shelter employees described dog behaviour correctly (using the canine behaviourists' opinion as a benchmark), which concerned videos showing *Reacts to people/dogs aggressive* and *Reacts to people/dogs non-aggressive* codes. While nearly 80% of shelter employees correctly identified the *Reacts to people/dogs non-aggressive* behaviour reported in **Paper II**, only approximately 50% of respondents correctly identified the *Reacts to people/dogs aggressive* codes. Instead, most individuals incorrectly recorded aggressive behaviour as non-aggressive. Thus, in **Paper I**, the true probability of aggression was probably underestimated. In **Paper II**, furthermore, the probability of higher category codes may have been reduced, meaning the probability of the most sociable *Friendly* code may have been inflated. Comparable estimates of validity are not available in the literature on shelter dogs. Overall, the identification of reactivity in dogs was accurate by shelter employees but employees were less accurate at identfying whether the motivation for the behaviour was aggressive or non-aggressive (e.g. frustrated).

Future research on dogs would benefit from more concerted use of confirmatory and reflective latent variable approaches. This could be fruitfully employed in meta-analyses of dog personality traits, where previous work has relied on the use of expert categorisation (e.g. Jones and Gosling, 2005; McGarrity *et al.*, 2015; Fratkin *et al.*, 2013). Whle the assumptions behind reflective models may be difficul to uphold, the benefit in employing this approach is the ability to test competing hypotheses and verify the robustness of the conclusions. In addition, behavioural ecologists have begun combining the application of latent variable models and hierarchical statistical models for studying personality and plasticity (Araya-Ajoy and Dingemanse, 2014; Martin and Suarez, 2017), which offers the chance to examine how a latent personality trait changes through time (e.g. behavioural plasticity of a trait). Indeed, human psychologists have been applying dynamic latent variable models for decades (e.g. Molenaar, 1985), techniques that have not been applied by animal personality researchers.

Network analysis is a promising and quickly-advancing area of research in human psychology, which can also handle time-varying and multi-level data structures. Bringmann *et al.* (2013), for example, introduced a multivariate vector-autoregressive model to analyse individual differences in depression symptom networks through time. These models combine the ability to analyse individual differences through time in a variety of statistical parameters (similar to the operational approach) with a network perspective on scientific constructs that sheds light onto the conceptual basis of personality and organisation of the behavioural phenotype more generally.

9 Conclusion

This thesis has examined the conceptual and methodological foundations of different approaches to studying personality and personality traits in dogs. In particular, the papers examined three broad approaches to studying personality in dogs, based on recent advances across ethology and psychology. The results demonstrate both strengths and weaknesses of the differing approaches.

The latent variable approach offers a powerful way of modelling the relationship between observed and unobserved variables. However, for the interpretation of the unobserved, latent variables to be clear, the choice of latent variable model requires careful thought. We demonstrated the utility of using confirmatory and reflective latent variable models for studying personality traits in dogs. Specifically, we found that, although the hypothesised latent variable model fit the data on inter-context aggressive behavior in shelters dogs well, key assumptions underlying the model (local independence and measurement invariance) were violated, implying that the aggressiveness towards people and dogs traits did not completely explain patterns of aggression in different contexts. Testing these assumptions in future research on the organisation of personality traits in dogs will be key in ensuring the robustness and reproducibility of the results.

By applying an operational approach, we found that shelter dogs varied in their degree of behavioural plasticity and predicitability over time, as well as in personality. Modelling predictability in shelter dogs substaintailly improved the predictive accuracy of the analyses, indicating that individual differences in within-individual behavioural variation is an integral component of dog behaviour that should be investigated in future work. At the same time, the amount of data on each individual dog over time at the shelter was relatively small, which meant that behavioural predictions entailed large uncertainty. This is a practical concern for shelters where the collection of large amounts of data on each individual is difficult, meaning methods that elucidate the amount of uncertainty in behavioural predictions will be important for informing realistic estimates of post-adoption behaviour.

Lastly, the network perspective offers a novel approach to understanding the organisation of the behavioural phenotype in animals, which encompasses the notion of personality. The application of network analysis to police patrol and detection dogs demonstrated a number of results supporting previous research on dog personality, as well as novel insights into the organisation of behaviour using centrality indices.

Given its flexibility and utility to a number of areas across science, the network approach may be the most promising approach for the study of behavioural phenotypes in animals, and could situate the study of personality within a more general scientific framework that is not hindered by criticisms of anthropomorphism.

References

Anusic, I and U Schimmack (2016). “Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations.” *Journal of Personality and Social Psychology* 110.5, pp. 766–781.

Araya-Ajoy, YG and NJ Dingemanse (2014). “Characterizing Behavioural ‘Characters’: An Evolutionary Framework”. *Proc. R. Soc. B* 281.1776, p. 20132645.

Asp, HE, WF Fikse, K Nilsson, and E Strandberg (2015). “Breed Differences in Everyday Behaviour of Dogs”. *Applied Animal Behaviour Science* 169, pp. 69–77.

Barnard, S, S Marshall-Pescini, A Pelosi, C Passalacqua, E Prato-Previde, and P Valsecchi (2017). “Breed, sex, and litter effects in 2-month old puppies’ behaviour in a standardised open-field test”. *Scientific Reports* 7.

Barnard, S, C Siracusa, I Reisner, P Valsecchi, and JA Serpell (2012). “Validity of model devices used to assess canine temperament in behavioral tests”. *Applied Animal Behaviour Science* 138.1, pp. 79–87.

Barnard, S *et al.* (2016). “Does Subjective Rating Reflect Behavioural Coding? Personality in 2 Month-Old Dog Puppies: An Open-Field Test and Adjective-Based Questionnaire”. *PLOS ONE* 11.3, e0149831.

Bar-Yam, Y (2016). “From Big Data To Important Information”. *arXiv:1604.00976 [nlin, physics:physics, q-bio, q-fin]*.

Beaujean, AA (2014). *Latent Variable Modeling Using R: A Step-by-Step Guide*. New York: Routledge.

Beekman, M and LA Jordan (2017). “Does the field of animal personality provide any new insights for behavioral ecology?” *Behavioral Ecology* 28.3, pp. 617–623.

Bell, AM, SJ Hankison, and KL Laskowski (2009). “The Repeatability of Behaviour: A Meta-Analysis”. *Animal Behaviour* 77.4, pp. 771–783.

Biro, PA and JA Stamps (2015). “Using Repeatability to Study Physiological and Behavioural Traits: Ignore Time-Related Change at Your Peril”. *Animal Behaviour* 105, pp. 223–230.

Boag, S (2015). “Personality assessment,‘construct validity’, and the significance of theory”. *Personality and Individual Differences* 84, pp. 36–44.

Bollen, K and R Lennox (1991). “Conventional Wisdom on Measurement: A Structural Equation Perspective”. *Psychological Bulletin* 110.2, pp. 305–314.

Borsboom, D (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Google-Books-ID: Fv3bAWvIDo4C. Cambridge University Press.

Borsboom, D (2006). “The Attack of the Psychometricians”. *Psychometrika* 71.3, pp. 425–440.

Braem, M, L Asher, S Furrer, I Lechner, H Würbel, and L Melotti (2017). “Development of the “Highly Sensitive Dog” questionnaire to evaluate the personality dimension “Sensory Processing Sensitivity” in dogs”. *PloS one* 12.5, e0177616.

Brandes, U (2001). “A faster algorithm for betweenness centrality”. *Journal of mathematical sociology* 25.2, pp. 163–177.

Bridgman, P (1954). “Remarks on the present state of operationalism”. *Scientific Monthly* 79.1, pp. 224–226.

Briffa, M and A Weiss (2010). “Animal personality”. *Current Biology* 20.21, R912–R914.

Bringmann, LF *et al.* (2013). “A network approach to psychopathology: new insights into clinical longitudinal data”. *PloS one* 8.4, e60188.

Brommer, JE (2013). “Variation in Plasticity of Personality Traits Implies That the Ranking of Personality Measures Changes between Environmental Contexts: Calculating the Cross-Environmental Correlation”. *Behavioral Ecology and Sociobiology* 67.10, pp. 1709–1718.

Budaev, SV (2010). “Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines”. *Ethology* 116.5, pp. 472–480.

Bullmore, E and O Sporns (2009). “Complex brain networks: graph theoretical analysis of structural and functional systems”. *Nature reviews. Neuroscience* 10.3, p. 186.

Capra, F and PL Luisi (2014). *The systems view of life: A unifying vision*. Cambridge University Press.

Carere, C and C Locurto (2011). “Interaction between animal personality and animal cognition”. *Current Zoology* 57.4, pp. 491–498.

Carpenter, B *et al.* (2016). “Stan: A Probabilistic Programming Language”. *J Stat Softw*.

Carter, AJ, WE Feeney, HH Marshall, G Cowlishaw, and R Heinsohn (2013). “Animal Personality: What Are Behavioural Ecologists Measuring?” *Biological Reviews* 88.2, pp. 465–475.

Casey, RA, B Loftus, C Bolster, GJ Richards, and EJ Blackwell (2014). “Human Directed Aggression in Domestic Dogs (Canis Familiaris): Occurrence in Different Contexts and Risk Factors”. *Applied Animal Behaviour Science* 152, pp. 52–63.

Cleasby, IR, 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”. *Methods in Ecology and Evolution* 6.1, pp. 27–37.

Cohen, AA, LB Martin, JC Wingfield, SR McWilliams, and JA Dunne (2012). “Physiological regulatory networks: ecological roles and evolutionary constraints”. *Trends in Ecology & Evolution* 27.8, pp. 428–435.

Cramer, AOJ *et al.* (2012). “Dimensions of Normal Personality as Networks in Search of Equilibrium: You Can’t Like Parties If You Don’t Like People”. *European Journal of Personality* 26.4, pp. 414–431.

Cramer, AOJ *et al.* (2016). “Major Depression as a Complex Dynamic System”. *PLOS ONE* 11.12, e0167490.

Cramer, AO (2012). “Why the item “23+ 1” is not in a depression questionnaire: Validity from a network perspective”. *Measurement: Interdisciplinary Research & Perspective* 10.1-2, pp. 50–54.

Crews, D (2013). “Review of Animal Personalities: Behavior, Physiology, and Evolution. Claudio Carere and Dario Maestripieri, editors.” *Integrative and Comparative Biology* 53 (5), pp. 873–875.

Croft, DP, R James, and J Krause (2008). *Exploring animal social networks*. Princeton University Press.

Dall, SR, AI Houston, and JM McNamara (2004). “The behavioural ecology of personality: consistent individual differences from an adaptive perspective”. *Ecology letters* 7.8, pp. 734–739.

David, M and SR Dall (2016). “Unravelling the philosophies underlying ‘animal personality’studies: A brief re-appraisal of the field”. *Ethology* 122.1, pp. 1–9.

Dingemanse, NJ and NA Dochtermann (2013). “Quantifying Individual Variation in Behaviour: Mixed-Effect Modelling Approaches”. *Journal of Animal Ecology* 82.1. DingDoch2013, pp. 39–54.

Dingemanse, NJ, AJN Kazem, D Réale, and J Wright (2010). “Behavioural Reaction Norms: Animal Personality Meets Individual Plasticity”. *Trends in Ecology & Evolution* 25.2, pp. 81–89.

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

Diverio, S *et al.* (2017). “Dogs’ coping styles and dog-handler relationships influence avalanche search team performance”. *Applied Animal Behaviour Science* 191, pp. 67–77.

Dochtermann, NA and SH Jenkins (2007). “Behavioural Syndromes in Merriam’s Kangaroo Rats (Dipodomys Merriami): A Test of Competing Hypotheses”. *Proceedings of the Royal Society of London B: Biological Sciences* 274.1623, pp. 2343–2349.

Dochtermann, NA and AB Nelson (2014). “Multiple facets of exploratory behavior in house crickets (Acheta domesticus): split personalities or simply different behaviors?” *Ethology* 120.11, pp. 1110–1117.

Duckworth, RA (2015). “Neuroendocrine Mechanisms Underlying Behavioral Stability: Implications for the Evolutionary Origin of Personality”. *Annals of the New York Academy of Sciences* 1360, pp. 54–74.

Duffy, DL, KA Kruger, and JA Serpell (2014). “Evaluation of a behavioral assessment tool for dogs relinquished to shelters”. *Preventive veterinary medicine* 117.3, pp. 601–609.

Edelen, MO and BB Reeve (2007). “Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement”. *Quality of Life Research* 16.1, p. 5.

Edelman, GM and JA Gally (2001). “Degeneracy and complexity in biological systems”. *Proceedings of the National Academy of Sciences* 98.24, pp. 13763–13768.

Epskamp, S, D Borsboom, and EI Fried (2016a). “Estimating Psychological Networks and their Accuracy: A Tutorial Paper”. *arXiv preprint*.

— (2017a). “Estimating psychological networks and their accuracy: A tutorial paper”. *Behavior Research Methods*.

Epskamp, S, AOJ Cramer, LJ Waldorp, VD Schmittmann, and D Borsboom (2012). “qgraph: Network Visualizations of Relationships in Psychometric Data”. *Journal of Statistical Software* 48.4, pp. 1–18.

Epskamp, S, J Kruis, and M Marsman (2017b). “Estimating psychopathological networks: Be careful what you wish for”. *PLOS ONE* 12.6, pp. 1–13.

Epskamp, S, GKJ Maris, LJ Waldorp, and D Borsboom (2016b). “Network Psychometrics”. *arXiv:1609.02818 [stat]*.

Fabrigar, LR, DT Wegener, RC MacCallum, and EJ Strahan (1999). “Evaluating the Use of Exploratory Factor Analysis in Psychological Research”. *Psychological Methods* 4.3, pp. 272–299.

Fadel, FR, P Driscoll, M Pilot, H Wright, H Zulch, and D Mills (2016). “Differences in Trait Impulsivity Indicate Diversification of Dog Breeds into Working and Show Lines”. *Scientific Reports* 6.

Foulley, JL and F Jaffrézic (2010). “Modelling and Estimating Heterogeneous Variances in Threshold Models for Ordinal Discrete Data via Winbugs/Openbugs”. *Computer Methods and Programs in Biomedicine* 97.1, pp. 19–27.

Foyer, P, N Bjällerhag, E Wilsson, and P Jensen (2014). “Behaviour and experiences of dogs during the first year of life predict the outcome in a later temperament test”. *Applied animal behaviour science* 155, pp. 93–100.

Fratkin, JL (2017). “Personality in Dogs”. *Personality in Nonhuman Animals*. Ed. by J Vonk, A Weiss, and SA Kuczaj. Cham: Springer International Publishing, pp. 205–224.

Fratkin, JL, DL Sinn, EA Patall, and SD Gosling (2013). “Personality Consistency in Dogs: A Meta-Analysis”. *PLOS ONE* 8.1, e54907.

Goddard, ME and RG Beilharz (1982). “Genetic and Environmental Factors Affecting the Suitability of Dogs as Guide Dogs for the Blind”. *Theoretical and Applied Genetics* 62.2, pp. 97–102.

Gosling, SD and OP John (1999). “Personality dimensions in nonhuman animals: a cross-species review”. *Current directions in psychological science* 8.3, pp. 69–75.

Green, CD (1992). “Of immortal mythological beasts: Operationism in psychology”. *Theory & Psychology* 2.3, pp. 291–320.

Harvey, ND, PJ Craigon, SA Blythe, GC England, and L Asher (2016a). “Social rearing environment influences dog behavioral development”. *Journal of Veterinary Behavior: Clinical Applications and Research* 16, pp. 13–21.

Harvey, ND, PJ Craigon, SA Blythe, GC England, and L Asher (2017). “An evidence-based decision assistance model for predicting training outcome in juvenile guide dogs”. *PloS one* 12.6, e0174261.

Harvey, ND *et al.* (2016b). “Test-retest reliability and predictive validity of a juvenile guide dog behavior test”. *Journal of Veterinary Behavior: Clinical Applications and Research* 11, pp. 65–76.

Honaker, J, G King, and M Blackwell (2015). *Amelia: A Program for Missing Data*.

Hoummady, S *et al.* (2016). “Relationships between personality of human–dog dyads and performances in working tasks”. *Applied Animal Behaviour Science* 177, pp. 42–51.

Hsu, Y and JA Serpell (2003). “Development and Validation of a Questionnaire for Measuring Behavior and Temperament Traits in Pet Dogs”. *Journal of the American Veterinary Medical Association* 223.9, pp. 1293–1300.

Hutton, EL (1945). “What is Meant by Personality?” *The British Journal of Psychiatry* 91.383, pp. 153–165.

Ilska, J *et al.* (2017). “Genetic Characterization of Dog Personality Traits”. *Genetics* 206.2, pp. 1101–1111.

Jones, AC and SD Gosling (2005). “Temperament and Personality in Dogs (Canis Familiaris): A Review and Evaluation of Past Research”. *Applied Animal Behaviour Science* 95.1, pp. 1–53.

Jones, AC (2008). *Development and Validation of a Dog Personality Questionnaire*. Google-Books-ID: HtSM2XwzMbcC. ProQuest.

Kauffman, SA (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA.

Kis, A, B Klausz, E Persa, Á Miklósi, and M Gácsi (2014). "Timing and Presence of an Attachment Person Affect Sensitivity of Aggression Tests in Shelter Dogs". *VETERINARY RECORD* 174.8, p. 196.

Kolaczyk, ED and G Csárdi (2014). *Statistical analysis of network data with R*. Vol. 65. Springer.

Koski, SE (2011). "How to Measure Animal Personality and Why Does It Matter? Integrating the Psychological and Biological Approaches to Animal Personality". *From Genes to Animal Behavior*. Ed. by M Inoue-Murayama, S Kawamura, and A Weiss. Primatology Monographs. Springer Japan, pp. 115–136.

Kruschke, J (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Google-Books-ID: FzvLAwAAQBAJ. Academic Press.

Ley, J, P Bennett, and G Coleman (2008). "Personality dimensions that emerge in companion canines". *Applied Animal Behaviour Science* 110.3, pp. 305–317.

Ley, JM, PC Bennett, and GJ Coleman (2009). "A refinement and validation of the Monash Canine Personality Questionnaire (MCPQ)". *Applied Animal Behaviour Science* 116.2, pp. 220–227.

Liddell, TM and JK Kruschke (2015). *Analyzing Ordinal Data: Support for a Bayesian Approach*. SSRN Scholarly Paper ID 2692323. Rochester, NY: Social Science Research Network.

Markus, KA and D Borsboom (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. 1 edition. New York, N.Y: Routledge.

Martin, JS and SA Suarez (2017). "Personality assessment and model comparison with behavioral data: A statistical framework and empirical demonstration with bonobos (Pan paniscus)". *American Journal of Primatology* 79.8. e22670, e22670–n/a.

Martin, JGA, DH Nussey, AJ Wilson, and D Réale (2011). "Measuring Individual Differences in Reaction Norms in Field and Experimental Studies: A Power Analysis of Random Regression Models". *Methods in Ecology and Evolution* 2.4, pp. 362–374.

Maul, A, DT Irribarra, and M Wilson (2016). "On the philosophical foundations of psychological measurement". *Measurement* 79, pp. 311–320.

McCrae, RR and PT Costa (1995). "Trait explanations in personality psychology". *European Journal of Personality* 9.4, pp. 231–252.

McGarrity, ME, DL Sinn, and SD Gosling (2015). "Which personality dimensions do puppy tests measure? A systematic procedure for categorizing behavioral assays". *Behavioural processes* 110, pp. 117–124.

McGarrity, ME, DL Sinn, SG Thomas, CN Marti, and SD Gosling (2016). “Comparing the predictive validity of behavioral codings and behavioral ratings in a working-dog breeding program”. *Applied Animal Behaviour Science* 179, pp. 82–94.

McGraw, KO and SP Wong (1996). “Forming inferences about some intraclass correlation coefficients.” *Psychological methods* 1.1, p. 30.

Mohan-Gibbons, H, E Weiss, and M Slater (2012). “Preliminary Investigation of Food Guarding Behavior in Shelter Dogs in the United States”. *Animals* 2.3, pp. 331–346.

Molenaar, PC (1985). “A dynamic factor model for the analysis of multivariate time series”. *Psychometrika* 50.2, pp. 181–202.

Mornement, KM, GJ Coleman, SR Toukhsati, and PC Bennett (2015). “Evaluation of the Predictive Validity of the Behavioural Assessment for Re-Homing K9’s (B.A.R.K.) Protocol and Owner Satisfaction with Adopted Dogs”. *Applied Animal Behaviour Science* 167, pp. 35–42.

Mornement, KM, GJ Coleman, S Toukhsati, and PC Bennett (2014). “Development of the behavioural assessment for re-homing K9’s (BARK) protocol”. *Applied Animal Behaviour Science* 151, pp. 75–83.

Nagasawa, M, S Kanbayashi, K Mogi, JA Serpell, and T Kikusui (2016). “Comparison of behavioral characteristics of dogs in the United States and Japan”. *Journal of veterinary medical science* 78.2, pp. 231–238.

Nakagawa, S and H Schielzeth (2010). “Repeatability for Gaussian and Non-Gaussian Data: A Practical Guide for Biologists”. *Biological Reviews of the Cambridge Philosophical Society* 85.4, pp. 935–956.

Nussey, D, A Wilson, and J Brommer (2007). “The evolutionary ecology of individual phenotypic plasticity in wild populations”. *Journal of evolutionary biology* 20.3, pp. 831–844.

Olson, KR *et al.* (2015). “Inconsistent Identification of Pit Bull-Type Dogs by Shelter Staff”. *The Veterinary Journal* 206.2, pp. 197–202.

Orritt, R (2015). “Dog Bites: A Complex Public Health Issue”. *Veterinary Record* 176.25, pp. 640–641.

Owczarczak-Garstecka, SC and OH Burman (2016). “Can Sleep and Resting Behaviours Be Used as Indicators of Welfare in Shelter Dogs (Canis lupus familiaris)?” *PloS one* 11.10, e0163620.

Patronek, GJ and J Bradley (2016). “No Better than Flipping a Coin: Reconsidering Canine Behavior Evaluations in Animal Shelters”. *Journal of Veterinary Behavior: Clinical Applications and Research* 15, pp. 66–77.

Pearl, J (2009). *Causality*. Cambridge university press.

Persson, ME, D Wright, LSV Roth, P Batakis, and P Jensen (2016). “Genomic Regions Associated with Interspecies Communication in Dogs Contain Genes Related to Human Social Disorders”. *Scientific Reports* 6.

Pervin, LA and OP John (1999). *Handbook of personality: Theory and research*. Elsevier.

Poulsen, AH, AT Lisle, and CJC Phillips (2010). “An Evaluation of a Behaviour Assessment to Determine the Suitability of Shelter Dogs for Rehoming”. *Veterinary Medicine International* 2010, e523781.

Preacher, KJ and RC MacCallum (2003). “Repairing Tom Swift’s Electric Factor Analysis Machine”. *Understanding Statistics* 2.1, pp. 13–43.

Proulx, SR, DE Promislow, and PC Phillips (2005). “Network thinking in ecology and evolution”. *Trends in Ecology & Evolution* 20.6, pp. 345–353.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rayment, DJ, BD Groef, RA Peters, and LC Marston (2015). “Applied Personality Assessment in Domestic Dogs: Limitations and Caveats”. *Applied Animal Behaviour Science* 163, pp. 1–18.

Réale, D, SM Reader, D Sol, PT McDougall, and NJ Dingemanse (2007). “Integrating Animal Temperament within Ecology and Evolution”. *Biological Reviews* 82.2, pp. 291–318.

Reise, SP, KF Widaman, and RH Pugh (1993). “Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance”. *Psychological Bulletin* 114.3, pp. 552–566.

Riemer, S, DS Mills, and H Wright (2014a). “Impulsive for life? The nature of long-term impulsivity in domestic dogs”. *Animal Cognition* 17.3, pp. 815–819.

Riemer, S, C Müller, Z Virányi, L Huber, and F Range (2014b). “The Predictive Value of Early Behavioural Assessments in Pet Dogs – a Longitudinal Study from Neonates to Adults”. *PLOS ONE* 9.7, e101237.

— (2016). “Individual and Group Level Trajectories of Behavioural Development in Border Collies”. *Applied Animal Behaviour Science* 180, pp. 78–86.

Roche, DG, V Careau, and SA Binning (2016). “Demystifying animal ‘personality’(or not): why individual variation matters to experimental biologists”. *Journal of Experimental Biology* 219.24, pp. 3832–3843.

Rosseel, Y (2012). “lavaan: An R Package for Structural Equation Modeling”. *Journal of Statistical Software* 48.2, pp. 1–36.

Rubin, DB (1976). “Inference and missing data”. *Biometrika* 63.3, pp. 581–592.

Ruedin, D (2016). *Agrmt: Calculate Agreement or Consensus in Ordered Rating Scales*.

Schimmack, U (2010). “What Multi-Method Data Tell Us about Construct Validity”. *European Journal of Personality* 24.3, pp. 241–257.

Schmittmann, VD, AOJ Cramer, LJ Waldorp, S Epskamp, RA Kievit, and D Borsboom (2013). “Deconstructing the Construct: A Network Perspective on Psychological Phenomena”. *New Ideas in Psychology*. On defining and interpreting constructs: Ontological and epistemological constraints 31.1, pp. 43–53.

Scott, JP and JL Fuller (2012). *Genetics and the Social Behavior of the Dog*. University of Chicago Press.

Seifert, L, J Komar, D Araújo, and K Davids (2016). “Neurobiological degeneracy: a key property for functional adaptations of perception and action to constraints”. *Neuroscience & Biobehavioral Reviews* 69, pp. 159–165.

SemTools Contributors (2017). *semTools: Useful tools for structural equation modeling*. R package version 0.4-15.910.

Serpell, JA and DL Duffy (2016). “Aspects of Juvenile and Adolescent Environment Predict Aggression and Fear in 12-Month-Old Guide Dogs”. *Frontiers in Veterinary Science* 3.

Sih, A, A Bell, J Johnson, and R Ziemba (2004). “Behavioral Syndromes: An Integrative Overview”. *The Quarterly Review of Biology* 79.3, pp. 241–277.

Sinn, DL, SD Gosling, and S Hilliard (2010). “Personality and Performance in Military Working Dogs: Reliability and Predictive Validity of Behavioral Tests”. *Applied Animal Behaviour Science* 127.1–2, pp. 51–65.

Spearman, C (1904). “"General Intelligence," Objectively Determined and Measured”. *The American Journal of Psychology* 15.2, pp. 201–292.

Stan Development Team (2016). *Rstan: R Interface to Stan*.

Stephen, JM and RA Ledger (2005). “An Audit of Behavioral Indicators of Poor Welfare in Kenneled Dogs in the United Kingdom”. *Journal of Applied Animal Welfare Science* 8.2, pp. 79–95.

Stevenson-Hinde, J, R Stillwell-Barnes, and M Zunz (1980). “Subjective assessment of rhesus monkeys over four successive years”. *Primates* 21.1, pp. 66–82.

Sundman, AS, M Johnsson, D Wright, and P Jensen (2016). “Similar Recent Selection Criteria Associated with Different Behavioural Effects in Two Dog Breeds”. *Genes, Brain and Behavior* 15.8, pp. 750–756.

Svartberg, K (2002). “Shyness–boldness Predicts Performance in Working Dogs”. *Applied Animal Behaviour Science* 79.2, pp. 157–174.

Svartberg, K and B Forkman (2002). “Personality Traits in the Domestic Dog (Canis Familiaris)”. *Applied Animal Behaviour Science* 79.2, pp. 133–155.

Szánthó, F, Á Miklósi, and E Kubinyi (2017). “Is your dog empathic? Developing a dog emotional reactivity survey”. *PloS one* 12.2, e0170397.

Taylor, KD and DS Mills (2006). “The Development and Assessment of Temperament Tests for Adult Companion Dogs”. *Journal of Veterinary Behavior: Clinical Applications and Research* 1.3, pp. 94–108.

Torchiano, M (2016). *effsize: Efficient Effect Size Computation*. R package version 0.6.4.

Uher, J (2011). “Individual behavioral phenotypes: An integrative meta-theoretical framework. Why “behavioral syndromes” are not analogs of “personality””. *Developmental psychobiology* 53.6, pp. 521–548.

Valsecchi, P, S Barnard, C Stefanini, and S Normando (2011). “Temperament Test for Re-Homed Dogs Validated through Direct Behavioral Observation in Shelter and Home Environment”. *Journal of Veterinary Behavior: Clinical Applications and Research* 6.3, pp. 161–177.

van den Berg, SM, HCM Heuven, L van den Berg, DL Duffy, and JA Serpell (2010). “Evaluation of the C-BARQ as a Measure of Stranger-Directed Aggression in Three Common Dog Breeds”. *Applied Animal Behaviour Science* 124.3–4, pp. 136–141.

Van der Maas, HL, CV Dolan, RP Grasman, JM Wicherts, HM Huizenga, and ME Raijmakers (2006). “A dynamical model of general intelligence: the positive manifold of intelligence by mutualism.” *Psychological review* 113.4, p. 842.

Van der Maas, HL, KJ Kan, and D Borsboom (2014). “Intelligence is what the intelligence test measures. Seriously”. *Journal of Intelligence* 2.1, pp. 12–15.

Van Borkulo, C, L Boschloo, D Borsboom, BW Penninx, LJ Waldorp, and RA Schoevers (2015). “Association of symptom network structure with the course of depression”. *JAMA psychiatry* 72.12.

Velicer, WF and DN Jackson (1990). “Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure”. *Multivariate behavioral research* 25.1, pp. 1–28.

Voith, VL *et al.* (2013). “Comparison of Visual and DNA Breed Identification of Dogs and Inter-Observer Reliability”. *American Journal of Sociological Research* 3.2, pp. 17–29.

Watanabe, S (2010). “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory”. *Journal of Machine Learning Research* 11.Dec, pp. 3571–3594.

Wicherts, JM and CV Dolan (2010). “Measurement Invariance in Confirmatory Factor Analysis: An Illustration Using IQ Test Performance of Minorities”. *Educational Measurement: Issues and Practice* 29.3, pp. 39–47.

Wilsson, E and PE Sundgren (1998). “Behaviour Test for Eight-Week Old Puppies—heritabilities of Tested Behaviour Traits and Its Correspondence to Later Behaviour”. *Applied Animal Behaviour Science* 58.1–2, pp. 151–162.

Wright, HF, DS Mills, and PM Pollux (2012). “Behavioural and physiological correlates of impulsivity in the domestic dog (Canis familiaris)”. *Physiology & behavior* 105.3, pp. 676–682.

10 Papers in order

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

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, Faculty of Biosciences, 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 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 positively correlated personality traits underlying aggression across contexts: aggressiveness towards people and aggressiveness towards dogs (comparative fit index: 0.96; Tucker-Lewis index: 0.95; root mean square error of approximation: 0.03). Aggression across contexts was moderately repeatable (towards people: intraclass correlation coefficient (ICC) = 0.479; towards dogs: ICC = 0.303). However, certain contexts related to aggressiveness towards people (but not dogs) shared significant residual relationships 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 demonstrate that animals show relatively consistent between-individual differences in behaviour, and that the behavioural phenotype is organised hierarchically into broad behavioural dimensions or personality traits (e.g. sociability, aggressiveness or boldness) that further exhibit inter-correlations to form behavioural syndromes (e.g. boldness with aggression; [1–5]). To interpret the complexity inherent in behavioural phenotypes, personality traits and behavioural syndromes are frequently inferred using latent variable statistical models [6], which reduce two or more measured variables (the *manifest* variables) into one or more lower-dimensional variables (the *latent* variables), following work in human psychology [7–10].

Many animal personality studies use *formative* models, such as principal components analysis, that construct composite variables comprised of linear combinations of manifest variables. However, formative models impose only weak assumptions about the relationships between latent variables and manifest variables [6,11]. For instance, formative models do not require manifest variables to be correlated with one another or illustrate internal consistency [11]. Because behavioural variables comprising personality traits are expected to correlate with each other [4], the utility of formative models to revealing underlying personality traits has been criticised in both animals [12,13] and humans [10,11,14,15]. Instead, researchers are increasingly using *reflective* models, such as factor analysis, including confirmatory approaches such as structural equation modelling (see [1,16–18]). Reflective models regress measured behaviours on one or more latent variables, incorporating measurement error and possibilities to compare *a priori* competing hypotheses [1,16,19].

Whilst reflective models offer a powerful framework to examine the latent variable structure of animal behaviour [19], they impose certain assumptions 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 are *local independence* and *measurement invariance*. Local independence implies that manifest variables should be independent of each other conditional on the latent variables [20,21]. For example, given a continuous latent variable θ (e.g. boldness) 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 covariance between manifest variables. Measurement invariance 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 [21–25]. In the previous example, this means that the expected values of the manifest variables Y_1 and Y_2 should remain the same across different groups, π (e.g. sex or different populations), for any fixed value of the latent variable θ_x e.g. $E(Y_1 \mid \theta_x) = E(Y_1 \mid \theta_x, \pi)$. For studies of personality, violations of local independence or measurement invariance highlight instances where the personality traits do not completely explain variation in the manifest variables, which may lead to misleading conclusions about the differences between individuals as a function of trait scores [25–27].

The goal of this study was evaluate local independence and measurement invariance in behavioural data on domestic dogs (*Canis lupus familiaris*). Dog personality has been of scientific interest for decades [28–30], both to predict the behaviour of dogs at future time points [31] and to elucidate behavioural traits pertinent to dogs' domestication history [32–35]. Research on personality in dogs has led to different numbers and composition of hypothesised personality traits with little consensus on how such traits should be compared within and between studies [36–38]. Dog personality studies frequently involve collection of data on a wide range of behaviours and, as a result, latent variable models are popular to reduce behavioural data into personality traits or dimensions [29]. Importantly, the predictive value of personality assessments in dogs has been inconsistent [31,39–43], perhaps most prominently in shelter dog personality assessments (e.g. see [31] for a review). Assessments of aggression are of particular concern, where aggression has been divided into different aggressiveness traits, such as owner-, stranger-, dog- or animal-directed factors [29,37,44,45]. Improving inference about aggressiveness in dogs is important because dog bites are a serious public health concern [46], especially for animal shelters rehoming dogs to new owners, and aggressive behaviour is undesirable to many organisations using dogs for various working roles [47].

Evaluating local independence and measurement invariance could help refine applied personality assessments on dogs. Local independence may be violated in standardised test batteries (a common assessment method; [48–50]) because the sequential administration of different behavioural subtests means that how dogs responds to one sub-test may influence their subsequent behavioural responses, as well as the responses of the dog handlers [31]. Identifying local independence could, thus, highlight which sub-tests can be interpreted as providing independent information. Local independence is also relevant to the development and analysis of dog personality questionnaires completed by dog owners, because the order in which the questions are presented or redundancy in the content of questions can lead to dependencies between participant responses not explained by the questionnaire's intended focus on the dog's behaviour [51].

Scientists are also concerned with understanding personality differences in dogs across a variety of conditions, including ontogeny, age, sex, breed and neuter status (e.g. [37, 42, 52–54]). Evaluating measurement invariance in personality assessments would allow researchers to confirm whether differences between individuals or groups of individuals in personality assessments reflect credible differences in personality trait scores or whether additional, unaccounted for factors are driving the differences. While it may be unrealistic for measurement invariance to hold in all instances, it is important to establish whether it holds for personality traits across basic biological variables such as age and sex, which are generally applicable to dog populations undergoing personality assessment and have previously been found to show interactions with personality traits, including playfulness, sociability, curiosity and aggressiveness [33, 55]. However, apart from van den Berg *et al*. [18] who assessed measurement invariance across breed groups, no studies have confirmed measurement invariance or local independence for personality traits.

In this paper, we assessed local independence and measurement invariance of aggressiveness in shelter dogs using a large sample of data on inter-context aggressive behaviour. First, we decomposed observations of aggression towards people and dogs across contexts into separate aggressiveness traits. Secondly, we assessed whether aggression in different contexts remained associated beyond that explained by latent levels of aggressiveness, testing local independence. Thirdly, we investigated whether the probability of aggression in different contexts assumed to be underpinned by the same aggressiveness trait was measurement invariant with respect to sex and age groups.

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 (Table 1). 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. Although dogs were from a variety of heritages (including purebreds and mongrels), the analyses here did not explore breed differences because the accurate visual assessment of breed in dogs with unknown heritage has been refuted [56–58].

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[1] before arrival / neutered at shelter / not neutered	1218 / 1665 / 1502
Relinquished by owners / returned to shelter / strays	2892 / 260 / 1591

[1] 358 dogs had unknown neuter status

Shelter environment

The shelter was composed of three different UK rehoming centres: a high-throughput, urban centre based at Battersea, London with capacity for approximately 150-200 dogs; a semi-rural/rural centre based at Old Windsor with capacity for approximately 100-150 dogs; and a rural centre based at Brands Hatch with capacity for approximately 50 dogs. All dogs arrived in an intake area of their respective rehoming centre and, when considered suitable for adoption, were moved to a ‘rehoming’ area that was partially open to the public between 1000 h and 1600 h. All kennels were indoors. Kennels varied in size, but were usually approximately 4m x 2m and included either a shelf and bedding alcove area, or a more secluded bedding area at the back of the kennel (see [59] for more details). At different times throughout the day, dogs had access to indoor runs behind their kennels. In each kennel block area, dogs were cared for (e.g. fed, exercised, kennel cleaned) by a relatively stable group of staff members, allowing the development of familiarity with staff members and offering some predictability for dogs after arrival at the shelter. Although data on the number of dogs in each kennel were incomplete, in the majority of cases dogs were kennelled singly for safety reasons. The shelter mainly operated between 0800 h and 1700 h each day. All dogs were socialised with staff and/or volunteers each day (often multiple times) except on rare occasions when it was deemed unsafe to handle a dog (when training/behavioural modification proceeded without physical contact). Dogs were provided water ad libitum and fed commercial complete dry and/or wet tinned food twice daily (depending on recommendations by veterinary staff). Dogs received daily tactile, olfactory and/or auditory enrichment/variety (e.g. toys, essential oils, classical music, time in a quiet ‘chill-out’ room).

Data collection

In the observational assessment procedure, trained shelter employees recorded observations of dog behaviour in a variety of contexts as part of normal shelter procedures. Behavioural observations pertaining to each context were completed using an ethogram specific to that context and recorded in a custom computer system. Multiple observations could be completed each day, although we retained only one observation in each context per day (the least desirable behaviour on that day; see below). The ethogram code that best described a dog's behaviour in a particular context during an observation was recorded by selecting it from a series of drop-down boxes (one for each context). Although staff could also add additional information in character fields, a full analysis of those comments was beyond the scope of 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 excluded offsite contexts (which had separate observation categories) because these were less frequently recorded and offsite records were more likely to be completed using second-hand information (e.g. from volunteers taking the dog offsite). We focused on observations of aggression in nine core onsite contexts that were most frequently completed by trained staff members: 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. If both occurred at the same time, aggression towards people was recorded. Therefore, *In kennel* and *Out of kennel* were each divided to reflect aggression shown towards people and towards dogs only, respectively. This resulted in 11 aggression contexts (Table 2) used as manifest variables in structural equation models to investigate latent aggressiveness traits. The average number of days between successive observations across these contexts and across dogs was 3.27 (SD = 2.08), and dogs had an average of 9.77 (SD = 13.41) observations within each context (N = 416,860 observations in total across dogs, contexts and days). Observations were recorded in the category that best described the scenario. Nonetheless, certain contexts could occur closely in space and time, which were investigated for violations of local independence, as explained below.

Table 2. Behavioural observation contexts in which each dog's reactions were analysed for the presence or absence of aggression.

Context	Definition
Handling	Informal handling by people (e.g. stroking non-sensitive areas, touching the collar, fitting a harness or lead).
In kennel towards people	People approaching or walking past the kennel.
In kennel towards dogs	Dogs in neighbouring kennels or dogs walking past the kennel.
Interactions with familiar people	When outside the kennel and familiar people (interacted with at least once before) approach, make eye contact, speak to or attempt to make physical contact with the dog.
Interactions with unfamiliar people	When outside the kennel and unfamiliar people (never interacted with before) approach, make eye contact, speak to or attempt to make physical contact with the dog.
Out of kennel towards people	When around people outside the kennel who may be a long distance away and who make no attempt to engage with the dog.
Out of kennel towards dogs	When around dogs outside the kennel that may be a long distance away and that are not encouraged to interact with the focal dog.
Eating food	When eating food (e.g. from a food bowl, or toy filled with food) and people approach within close proximity or attempt to touch the food container.
Interactions with toys	When interacting with toys and people approach within close proximity or attempt to touch the toy.
Interactions with female dogs	During structured interaction 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	During structured interaction 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.

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 (Table S1). This was performed because the overall prevalence of aggression was low, with only 1.06% of all observations across days 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. We interpret aggressiveness here as a between-individual difference variable.

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 first video had an ethogram of 13 possible behaviours to choose from, and the second had 11 behaviours. The authors were blind to the selection of videos shown and to the video coding sessions with shelter employees.

Data analysis

All data analysis was conducted in R version 3.3.2 [60].

Validity of behaviour recordings

The degree to which shelter employees could recognise and correctly record aggressive behaviour from the videos (chosen by experienced behaviourists at the shelter) was determined by the percentage of participants who correctly identified the 2 videos as showing examples of aggression.

Missing data

Data were missing when dogs did not experience particular contexts while at the shelter. The missing data rate was between 0.06% and 5% for each context, except for the *Interactions with female dogs* and *Interactions with male dogs* categories which had 17% and 18% of missing values, respectively (because structured interactions with other dogs did not arise as frequently). Moreover, 16% and 8% of dogs were missing weight measurement and neuter status data, respectively, which were independent variables statistically controlled for in subsequent analyses. We created 5 multiply imputed data sets (using the *Amelia* package; [61]), upon which all following analyses in the sections below were conducted and results pooled. The multiple imputation took into account the hierarchical structure of the data (observations within dogs), all independent variables reported below, and the data types (ordered binary variables for the context data, positive-continuous for weight measurements, nominal for neuter status; see the R script). 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. Since positive correlations between different aggressiveness traits have been reported in dogs [55], we compared a model where the latent variables were orthogonal to a model where variables were allowed to covary. Models were fit using the *lavaan* package [62], with the weighted least squares mean and variance adjusted (WLSMV) estimator and theta/conditional parameterisation, as recommended for categorical dependent variables [8,63,64]. 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 [65]. 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 test statistics and 95% confidence intervals (CI).

Local independence

We tested the assumption of local independence by re-fitting the best-fitting structural equation model with residual covariances specified between context variables. To maintain a theoretically driven approach (see [66] regarding the best practice of including residual covariances in structural equation models) and model identifiability, we only tested a predefined set of covariances based on which contexts shared close temporal-spatial relationships. First, we allowed covariances between *Handling* with *In kennel towards people*, *Interactions with familiar people*, *Interactions with unfamiliar people* and *Interactions with toys*, respectively, since the *Handling* context could directly succeed these other contexts. The residual covariance 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 covariance 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 covariances between *Out of kennel towards people*, *Interactions with familiar* *people* and *Interactions with unfamiliar people*. Similarly, and lastly, we estimated the three-way covariances between *Out of kennel towards dogs, Interactions with female dogs* and *Interactions with male dogs*. No covariances were inspected between *In kennel towards dogs* and other aggressiveness towards dogs contexts since large time gaps were more likely to separate observations between those contexts.

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. These models were analogous to the 1-parameter item response theory model, which 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. This model can be expressed as a hierarchical logistic regression model [67,68], 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 (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 that the intercept represented the average probability of aggression across the levels of each categorical predictor. 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; [52]), 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 [69].

Computation

Models were computed using the probabilistic programming language Stan version 2.15.1 [70], using Hamiltonian Monte Carlo, a type of Markov Chain Monte Carlo (MCMC) algorithm, to sample from the posterior distribution. 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-specific intercept parameters was given a half-Cauchy prior distribution with mean 0 and shape 2. Each model was run with 4 chains of 2,000 iterations with a 1,000 step warm-up period. The Gelman-Rubin statistic (ideally < 1.05) and visual assessment of traceplots were used to assess MCMC convergence. We checked the accuracy of the model predictions against the raw data using graphical posterior predictive checks. For plotting purposes, predicted probabilities of aggression were obtained by marginalising over the random effects (explained in the Supporting

Information). Regression coefficients were expressed as odds ratios and were summarised by their mean and 95% Bayesian highest density interval (HDI), representing the 95% most probable parameter values. To compare levels of categorical variables and their interactions, we computed the 95% HDI of the differences between the respective posterior distributions.

Model selection & parameter inference

Models were run on each imputed data set and their respective posterior distributions were averaged to attain a single posterior distribution for inference. Adopting a Bayesian approach allowed the estimation of interaction parameters (i.e. testing measurement invariance) without requiring corrections for multiple comparisons as in null hypothesis significance testing [71]. Nonetheless, models included a large number of estimated parameters. Two strategies were employed to guard against over-fitting of models to data. First, we selected the model with the best out-of-sample predictive accuracy given the number of parameters based on the Widely Applicable Information Criterion (WAIC; using the R package *loo* [72]). Four variants of each model were computed: two-way interactions between contexts and age and contexts and sex, respectively (model 1), a single interaction with sex but not with age (model 2), a single interaction with age but not with sex (model 3), and no interactions (model 4). All models included the mediating independent variables above. Second, to avoid testing point-estimate null hypotheses, the effect of a parameter was only considered credibly different from zero if the odds ratio exceeded the region of practical equivalence (ROPE; see [73]) around an odds ratio of 1 from 0.80 to 1.25. An odds ratio of 0.80 or 1.25 indicates a 20% decrease or increase (i.e. 4/5 or 5/4 odds), respectively, in the odds of an outcome, frequently used in areas of bioequivalence testing (e.g. [74]), which we here considered to be small enough to demonstrate a negligible effect in the absence of additional information. If a 95% HDI fell completely within the ROPE, the null hypothesis of no credible influence of that parameter was accepted; if a 95% HDI included part of the ROPE, then the parameter's influence was left undecided [73].

Ethics statement

Permission to use and publish the data was received from the shelter. Approval from an ethical review board was not required for this study.

Data accessibility

Supporting Information (data, R script, Stan model code, Tables S1-4) can be found at: https://github.com/ConorGoold/GooldNewberry_aggression_shelter_dogs.

Results

Validity of behaviour recordings

For the video showing aggression towards people, 52% of participants identified the behaviour correctly as aggression and 42% identified the behaviour as non-aggressive but (similarly) reactive behaviour (see definitions above). For the video showing aggression towards dogs, 53% identified the behaviour correctly and 44% identified the behaviour as non-aggressive but reactive behaviour. For the 12 other videos not showing aggression, only 1 person incorrectly coded a video as aggression towards people and 3 people incorrectly coded videos as aggression towards dogs.

Structural equation models

The raw tetrachoric correlations between the aggression contexts were all positive, particularly between contexts recording aggression towards people and dogs, respectively, supporting their convergent validity (Table S2). The model with correlated latent variables fit marginally better (CFI: 0.96; TLI: 0.95; RMSEA: 0.03) than the model with uncorrelated variables (CFI: 0.94; TLI: 0.92; RMSEA: 0.04). All regression coefficients of the model with correlated latent variables were positive and significant (i.e. the 95% CI did not include zero), and the latent variables shared a significant positive covariance (Table 3).

Table 3. Parameter estimates from the best-fitting structural equation model.

Parameter	**Estimate**	**SE**	***t* value**	**95% CI**
Handling[a]	0.81	0.06	14.25	[0.70, 0.92]
In kennel towards people[a]	1.29	0.09	14.17	[1.12, 1.46]
Out of kennel towards people[a]	0.83	0.07	11.99	[0.69, 0.96]
Interactions with familiar people[a]	0.96	0.07	14.23	[0.83, 1.09]
Interactions with unfamiliar people[a]	1.54	0.12	12.46	[1.23, 1.78]
Eating food[a]	0.70	0.06	12.33	[0.59, 0.81]
Interactions with toys[a]	0.51	0.06	8.32	[0.39, 0.63]
In kennel towards dogs[b]	0.70	0.06	11.94	[0.59, 0.82]
Out of kennel towards dogs[b]	0.47	0.04	10.80	[0.38, 0.55]
Interactions with female dogs[b]	0.87	0.07	12.05	[0.72, 1.02]
Interactions with male dogs[b]	0.88	0.07	12.23	[0.74, 1.03]
Covariance: People ~ Dogs	0.26	0.03	7.94	[0.19, 0.33]

[a] Contexts reflecting aggressiveness towards people

[b] Contexts reflecting aggressiveness towards dogs

Local independence

Allowing the pre-defined residuals to co-vary in the best-fitting structural equation model resulted in a better fit (CFI = 0.98; TLI = 0.97; RMSEA: 0.03). Significant negative covariances were observed between the *Handling* and *In kennel towards people* contexts

(Table 4) and the *Handling* and *Interactions with unfamiliar people* contexts. A significant positive covariance was observed between *Out of kennel towards people* and *Interactions with unfamiliar people* contexts. No significant residual covariances between contexts reflecting aggressiveness towards dogs were observed.

Table 4. Estimated residual covariances between contexts.

Residual covariances	**Estimate**	**SE**	***t* value**	**95% CI**
Handling ~ In kennel towards people[a]	-0.60	0.21	-2.86	[-1.01, -0.19]
Handling ~ Interactions with familiar people[a]	0.16	0.09	1.84	[-0.01, 0.33]
Handling ~ Interactions with unfamiliar people[a]	-0.48	0.19	-2.49	[-0.86, -0.10]
Handling ~ Interactions with toys[a]	0.14	0.07	1.85	[-0.01, 0.28]
Out of kennel towards people ~ Interactions with familiar people[a]	0.04	0.08	0.49	[-0.12, 0.20]
Out of kennel towards people ~ Interactions with unfamiliar people[a]	0.24	0.09	2.56	[0.06, 0.42]
Interactions with familiar people ~ Interactions with unfamiliar people[a]	-0.02	0.12	-0.16	[-0.25, 0.21]
Out of kennel towards dogs ~ Interactions with female dogs[b]	-0.55	0.48	-1.15	[-1.50, 0.40]
Out of kennel towards dogs ~ Interactions with male dogs[b]	-0.45	0.40	-1.13	[-1.22, 0.33]
Interactions with female dogs ~ Interactions with male dogs[b]	-0.24	0.50	-0.49	[-1.23, 0.74]

[a] Contexts reflecting aggressiveness towards people

[b] Contexts reflecting aggressiveness towards dogs

441

Measurement invariance

Separate models were run for contexts reflecting aggressiveness towards people and aggressiveness towards dogs. All models converged. Posterior predictive checks of model estimates reflected the raw data (Figs 1 and 2). The full measurement invariance model (model 1) including interactions between contexts and sex and contexts and age groups had the best out-of-sample predictive accuracy for both the aggressiveness towards people and aggressiveness towards dogs models, respectively, illustrated by the lowest WAIC values (Table 5). Since some models included numerous interactions, we provide an overall summary of the main results below (Figs 1 and 2) with full parameter estimates provided in Tables S3 and S4.

Table 5. Mean ± standard error of the Widely Applicable Information Criteria (WAIC) values (lower is better) per model and aggressiveness variable.

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

Aggressiveness towards people

The odds of aggression towards people, across categorical predictors and for an average dog of mean weight and length of stay at the shelter, were 0.022 (HDI: 0.021 to 0.024), a probability of approximately 2%. On average, aggression was most likely in the *In kennel towards people* context (OR = 0.054; HDI: 0.049 to 0.058) and least probable in the *Interactions with toys* context (OR = 0.008; HDI: 0.007 to 0.009).

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

Apart from lower odds of aggression in 4 to 10 month olds compared to 10 month to 3 year old dogs (OR = 0.638; HDI: 0.565 to 0.705), there was no simple influence of age group on aggressiveness. Between the 4 to 10 months old and 3 to 6 years old groups, 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; HDI: 0.400 to 0.642) and *Eating food* and *Interactions with unfamiliar people* (OR = 1.721; HDI: 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; HDI: 0.355 to 0.606) and *Out of kennel towards people* and *Interactions with unfamiliar people* (OR = 2.051; HDI: 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; HDI: 0.409 to 0.631) and *Out of kennel towards people* and *Interactions with unfamiliar people* (OR = 2.349; HDI: 1.891 to 2.925), and smaller differences between *Eating food* and *Interactions with familiar people* (OR = 0.576; HDI: 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* contexts compared to the other age groups (Fig 1B; Table S3). Aggression when *Eating food* and in *Interactions with toys* contexts also increased compared to that expressed by younger dogs, resulting in credible differences between, for instance, *Eating food* and *Interactions with familiar people* contexts between dogs aged 10 months to 3 years and over 6 years (OR = 0.379; HDI: 0.300 to 0.465) and between *Out of kennel towards people* and *Interactions with toys* contexts between over 6 year olds and all other age groups (Table S3).

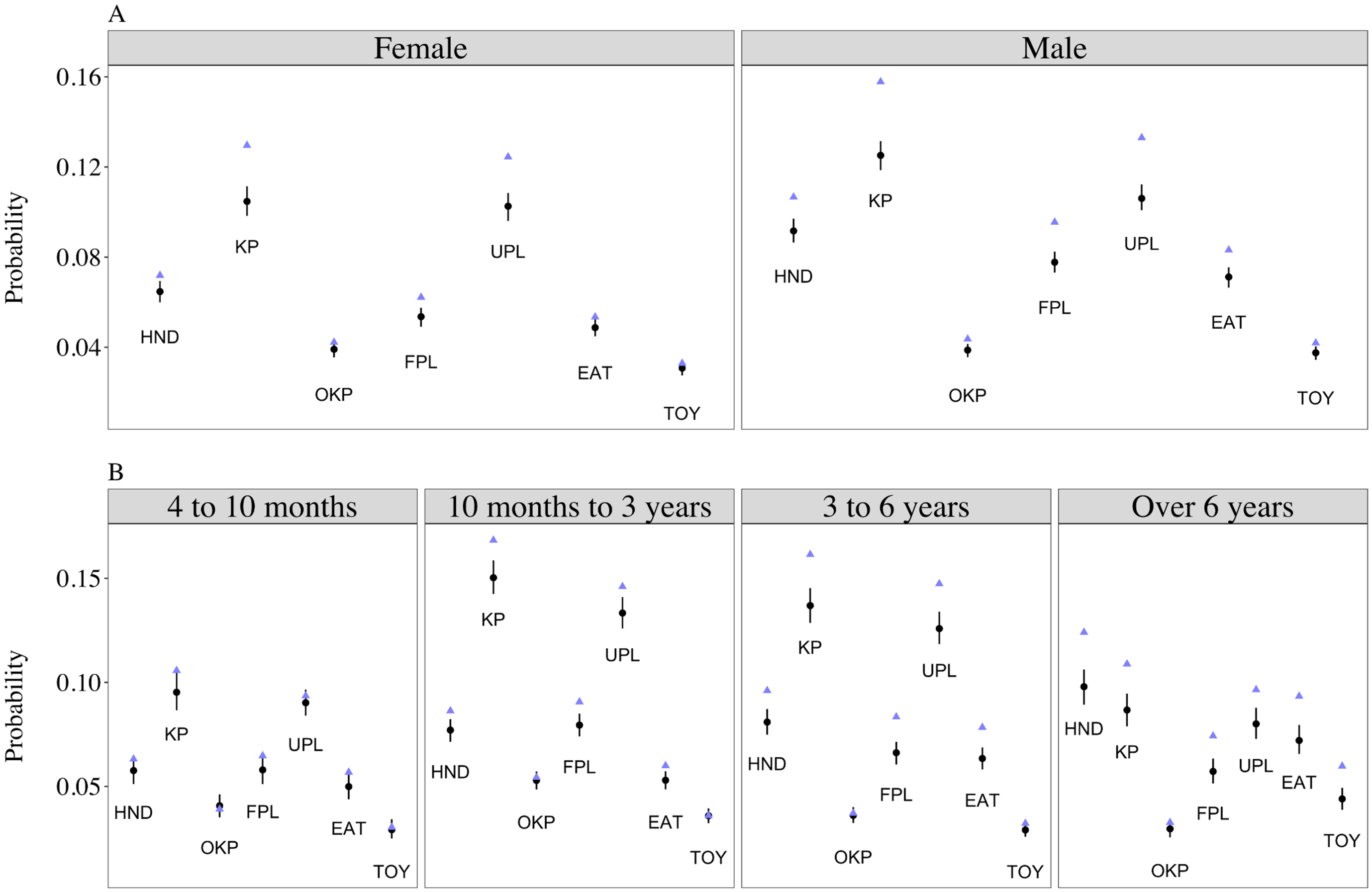


Fig 1. Predicted probabilities of aggression towards people in different contexts by sex (panel A) and age groups (panel B). Black points and vertical lines show mean and 95% highest density intervals of model parameter estimates; blue triangles show raw sample data. Model estimates were obtained by marginalising over the random effects (see the Supporting Information). 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*).

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 (HDI: 0.168 to 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; HDI: 0.198 to 0.217) and least likely in the *In kennel towards dogs* context (OR = 0.099; HDI: 0.094 to 0.104; Fig 2; Table S4).

No credible mean-level differences existed between females and males (OR = 1.187; HDI: 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; HDI: 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; HDI: 0.590 to 0.732) and *In kennel towards dogs* and *Out of kennel towards dogs* (OR = 1.420; HDI: 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; HDI: 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 *Interactions with male dogs* contexts tended to increase relative to other contexts. For instance, the relationship between *Interactions with female dogs* and *Out of kennel towards dogs* contexts reversed in direction between 4 to 10 month and 10 month to 3 year olds (OR = 0.595; HDI: 0.495 to 0.688) as did the relationship between *Interactions with male dogs* and *Out of kennel towards dogs* contexts (OR = 0.499; HDI: 0.422 to 0.575). The relationship between *In kennel towards dogs* and *Out of kennel towards dogs* contexts also changed across age groups (Fig 2B; Table S4). Four to 10 months old were more likely to show aggression in *Out of kennel towards dogs* than *In kennel towards dogs* contexts, but the difference was smaller in 10 months to 3 year olds (OR = 0.608; HDI: 0.505 to 0.728) and in over 6 year olds (OR = 0.396; HDI: 0.316 to 0.481). The latter relationship was reversed in 3 to 6 year olds compared to 4 to 10 month old dogs (OR = 0.277; HDI: 0.227 to 0.331) and 10 month to 3 year old dogs (OR = 0.456; HDI: 0.396 to 0.516).

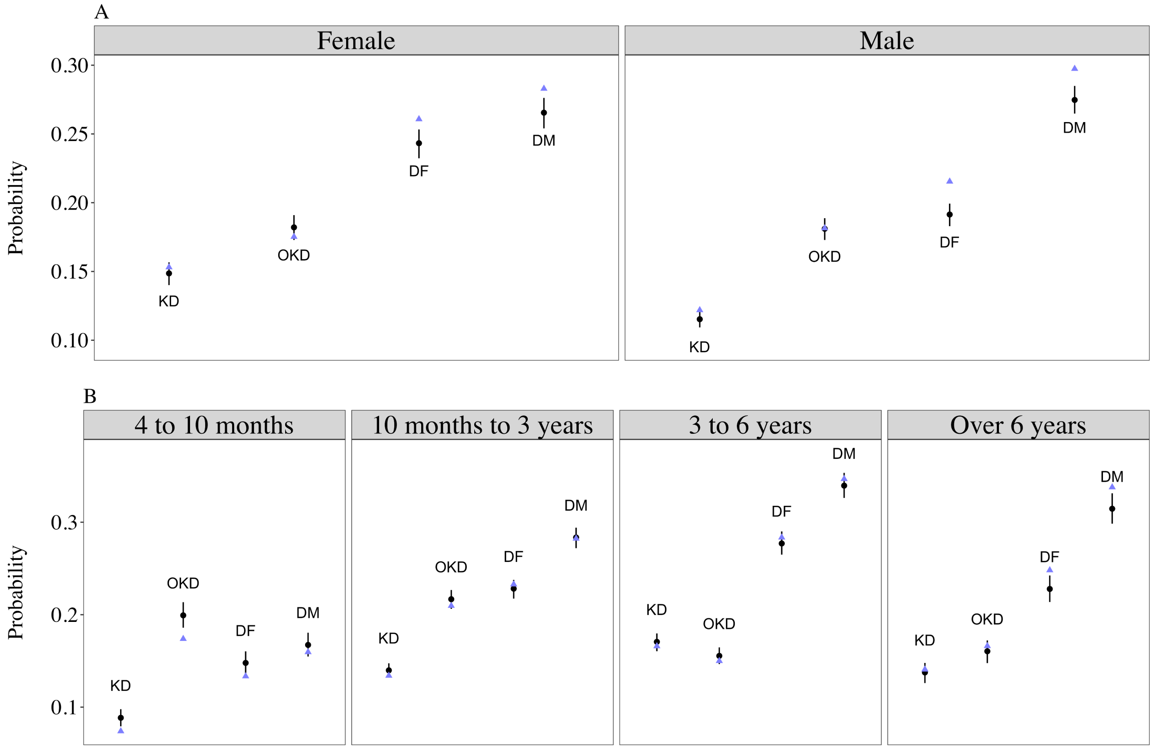


Fig 2. Predicted probabilities of aggression towards dogs in different contexts by sex (panel A) and age groups (panel B). Black points and vertical lines show mean and 95% highest density intervals of model parameter estimates; blue triangles show raw sample data. Model estimates were obtained by marginalising over the random effects (see the Supporting Information). 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*).

Repeatability

Both aggressiveness towards people and dogs showed moderate repeatability across contexts (ICC_{people} = 0.479; HDI: 0.466 to 0.491; ICC_{dogs} = 0.303; HDI: 0.291 to

0.315), although aggressiveness towards people was more repeatable than aggressiveness towards dogs ($ICC_{difference}$ = 0.176; HDI: 0.158 to 0.192).

Discussion

In this study, we have examined local independence and measurement invariance of 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 concept of animal personality, which is used to describe behaviour that shows moderately consistent between-individual differences across time or contexts, and is characterised by multiple observed behaviours being decomposed into lower-dimensional behavioural traits [4]. However, we found violations of local independence between contexts with close temporal-spatial relationships and measurement invariance with respect to sex and age groups, highlighting potential measurement biases.

Local independence implies that the association between manifest variables is greater than that explained by the latent variable. For aggressiveness towards people, aggression in the *Handling* context was negatively related with the *In kennel towards people* and *Interactions with unfamiliar people* contexts, while positive covariances were present between *Out of kennel towards people* and *Interactions with unfamiliar people* contexts. Violations of local independence may arise through shared method variance [75–78] or unmodelled latent variables influencing manifest variables [79,80]. If a dog showed aggression when an unfamiliar person approached, it may be less likely to be handled by that person, which may explain the negative residual covariations between the *Handling* and *In kennel towards people* and *Interactions with unfamiliar people* contexts, respectively. These contexts were, in fact, positively correlated when latent levels of aggressiveness were not accounted for (Table S4). In addition, the positive residual correlation between *Out of kennel towards people* and *Interactions with unfamiliar people* may be mediated by additional traits of interest to personality researchers, such as fearfulness or anxiety [29,81], if dogs who are fearful of interacting with unfamiliar people are more likely to show aggression beyond that described by a latent aggressiveness trait.

While authors have argued that greater standardisation and validation of personality assessments is key to ensuring the accurate measurement of underlying traits [36,48,49], it may be untenable to avoid dependencies between testing contexts. Displays of aggression in one sub-test will likely change how people conduct future sub-tests with the same dog, regardless of test standardisation. Human psychologists have argued that violations of local independence are a natural consequence of the organisation of behaviour as a complex dynamic system [82,83], which unfolds with respect to time- and context-dependent constraints [84]. Thus, awareness of local independence and its violation could facilitate closer understanding of the dynamics driving personality test responses beyond explanations purely based on personality traits.

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]. We found that the probability of aggression across contexts was dependent on sex and age conditional on latent levels of aggressiveness (Figs 1 and 2; Tables S3 and S4). 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). As with local independence, different behavioural variables unaccounted for in this study may result in violations of measurement invariance. 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. 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. These results suggest that older dogs in shelter populations may be less tolerant during close interactions with people (i.e. handling, people in the vicinity of their food and toys) compared to other contexts, which may driven by other quantifiable factors such as pain or sensitivity (e.g. [29]).

Although we have identified violations of both local independence and measurement invariance, we remain cautious about hypothesising *a posteriori* about their causes. Personality traits in animal behaviour are typically defined operationally, based on the statistical repeatability of quantifiable behaviour [77,85,86]. As discussed in human personality psychology, operational definitions can be ontologically ambiguous [87,88]. That is, while operational definitions facilitate experimentation in animal personality [4], they do not necessarily designate biological mechanisms underlying trait expression. For example, Budaev and Brown remark that boldness, defined as a propensity to take risks, could encompass a range of distinct personality traits, each with a different biological basis [75]. Whilst reflective latent variable models allow researchers to test hypotheses about the relatedness of measured behaviours via one or more underlying traits, they have also been criticised as ambiguous [82]. For example, it is uncertain what reflective latent variables may represent in biological organisation [87] or even whether they are features individuals possess or simply emergent features of between-individual differences [89,90]. Such considerations highlight the importance of research on the proximate mechanisms of personality [85] and longitudinal data analyses to separate between- from within-individual behavioural variation [91,92].

A number of authors have emphasised the poor predictive value of aggression tests in shelter dogs [39–41,50] and that low occurrence of aggression specifically can make its accurate measurement difficult [40]. The probability of observing aggression on any particular day was low in this study (approximately 1%), and the number of dogs who, on average, showed aggression to people at least once while at the shelter was much lower than the number that showed aggression towards dogs, on average (Figs 1 and 2). Nonetheless, our evaluations of validity indicated that between 40 and 45% of the shelter employees mistook observations of aggression for non-aggressive responses (e.g. over-excitement and frustration when seeing people/dogs), meaning that the true probability of aggression was potentially under-estimated (although incorrectly coding other behaviours as aggression also occurred, albeit rarely). Moreover, our assessments of validity were based on shelter staff evaluations of brief video recordings that may be less reliable than the live, spontaneous behavioural recordings upon which our main analyses were based, resulting in a lower percentage of correctly identified instances of aggression. For the two videos being evaluated, the shelter employees had 13 and 11 different behavioural codes, respectively, to choose from to describe the behaviours observed. Thus, while employees as a whole were undecided about whether the motivation for the behaviour was aggressive or non-aggressive, the vast majority of employees described the behaviour as reactive, despite potentially erring on the side of caution by labelling aggressive behaviours as non-aggressive. Comparable estimates of validity are not present in the literature on dog personality, but are particularly important in shelter settings where accurate recording of aggression is paramount. It is also worth noting that how to assess validity has received much debate (e.g. [87,92]). In this study, we used expert judgement as a benchmark to which shelter employees' responses were compared, but validity is frequently assessed in dog personality by inspecting patterns of correlation coefficients between similar and dissimilar behaviours (e.g. convergent or divergent validity; [29]). This is less directly interpretable than reporting the percentage of answers that were correct, as used here. Moreover, the predictive validity of personality assessments in dogs have been inconsistent (e.g. [40-42]). More discussion of the concept of validity, and how best to assess it, is warranted in studies of dog personality.

Infrequent occurrence and/or recording of aggression may also limit accurate predictions of future behaviour. Patronek and Bradley [50] 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 [31,40,50]. 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 covariances remained between certain behavioural contexts related to aggressiveness towards people, violating the assumption of local independence. In addition, aggression in different contexts showed differential patterns of response across sex and age, indicating a lack of measurement invariance. Violations of local independence and measurement invariance imply that the aggressiveness towards people and dogs traits did not completely explain patterns of aggression in different contexts, or that inferences based on these hypothesised personality traits may in fact be misleading. We encourage researchers to more closely assess the measurement assumptions underlying reflective latent variable models before making conclusions about the effects of, or factors influencing, personality.

Acknowledgements

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

References

1. Dochtermann NA, Jenkins SH. Behavioural syndromes in Merriam's kangaroo rats (*Dipodomys merriami*): a test of competing hypotheses. Proc R Soc B. 2007;274: 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. Budaev DS. How many dimensions are needed to describe temperament in animals: a factor reanalysis of two data sets. Int J Comp Psychol. http://cogprints.org/5478/; 1998.

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

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

9. 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

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

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

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

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

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

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

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

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

18. van den Berg SM, Heuven HCM, van den Berg L, Duffy DL, Serpell JA. Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog

breeds. Appl Anim Behav Sc. 2010;124: 136–141. doi:10.1016/j.applanim.2010.02.005

19. Martin JS, Suarez SA. Personality assessment and model comparison with behavioral data: a statistical framework and empirical demonstration with bonobos (*Pan paniscus*). Am J Primatol. 2017;9999: e22670. doi: 10.1002/ajp.22670

20. Bartholomew DJ. Factor analysis for categorical data. J R Stat Soc Series B Stat Methodol. 1980;42: 293–321. http://www.jstor.org/stable/2985165

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

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

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

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

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

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

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

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

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

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

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

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

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

34. Barnard S, Marshall-Pescini S, Pelosi A, Passalacqua C, Prato-Previde E, Valsecchi P. Breed, sex, and litter effects in 2-month old puppies' behaviour in a standardised open-field test. Sci Rep. 2017;7: 1 – 8. doi:10.1038/s41598-017-01992-x.

35. Svartberg K. Breed-typical behaviour in dogs – historical remnants or recent constructs? Appl Anim Behav Sci. 2006;96: 293 – 313. doi: 10.1016/j.applanim.2005.06.014

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

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

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

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

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

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

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

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

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

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

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

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

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

49. Haverbeke A, Plujimakers J, Diederich C. Behavioral evaluations of shelter dogs: literature review, perspectives, and follow-up within the European member state's legislation with emphasis on the Belgian situation. J Vet Behav. 2015;10: 5 – 11. doi: 10.1016/j.jveb.2014.07.004

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

51. Edelen MOE, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16: 5 – 18. doi: 10.1007/s11136-007-9198-0

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

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

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

55. 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

56. 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. doi: 10.5923/j.sociology.20130302.02

57. Olson KR, Levy JK, Norby B, Crandall MM, Broadhurst JE, Jacks S, Barton RC, Zimmerman MS. Inconsistent identification of pit bill-type dogs by shelter staff. Vet J. 2015;206: 197 – 202. doi:10.1016/j.tvjl.2015.07.019

58. Simpson RJ, Simpson K, VanKavage L. Rethinking dog breed identification in veterinary practice. J Am Vet Med Assoc. 2012;241: 1163 – 1166. doi: 10.2460/javma.241.9.1163

59. 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

60. 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/

61. 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

62. 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/

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

64. 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

65. 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/

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

67. 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

68. 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.

69. 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

70. Stan Development Team. Stan modeling language users guide and reference manual. 2016. Version 2.15.0. https://github.com/stan-dev/stan/releases/download/v2.14.0/stan-reference-2.14.0.pdf

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

72. 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

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

74. 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

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

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

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

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

79. Cole DA, Ciesla JA, Steiger JH. The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. Psychol Methods. 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. 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

86. 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

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

88. 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

89. 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

90. 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

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

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

93. Borsboom D, Cramer AO, Kievit RA, Zand Scholten A, Franic S. The end of construct validity. In Lissitz, R, editor. The concept of validity. Information Age Publishers; 2010. pp. 135–170.

Supporting Information

Table S1. Counts of aggression per context. The number of dogs who had 0, 1, and > 1 observations of aggression while at the shelter.

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

Table S4. Bayesian hierarchical model parameter estimates for aggression towards dogs in different contexts. Mean and 95% highest density interval (HDI) 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% HDI falls completely outside the region of practical equivalence (ROPE); NULL = 95% HDI falls completely inside the ROPE; ROPE = 95% HDI partly covers the ROPE.

Modelling personality, plasticity and predictability in shelter dogs

Modelling personality, plasticity and predictability in shelter dogs

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

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

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

Abstract

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

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

Introduction

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

Modest repeatability of behaviour has been documented in domestic dogs (*Canis lupus familiaris*), providing evidence for personality variation. For instance, using meta-analysis, Fratkin *et al.* [13] found an average Pearson's correlation of behaviour through time of 0.43, explaining 19% of the behavioural variance between successive time points (where the average time interval between measurements was 21 weeks). However, the goal of personality assessments in dogs is often to predict an individual dog's future behaviour (e.g. working dogs: [14,15]; pet dogs: [16]) and, thus, it is important not to confuse the stability of an individual's behaviour relative to the behaviour of others with stability of intra-individual behaviour. That is, individuals could vary their behaviour in meaningful ways in response to internal (e.g. ontogeny) and external (e.g. environmental) factors while maintaining differences from other individuals. When time-related change in dog behaviour has been taken into account, behavioural change at the group-level has been of primary focus (e.g. [16–18]) and no studies have explored the heterogeneity of residual variance within each dog. The predominant focus on inter-individual differences and group-level patterns of behavioural change risks obscuring important individual-level heterogeneity and may partly explain why a number of dog personality assessment tools have been unreliable in predicting future behaviour [14–16,19].

Of particular concern is the low predictive value of shelter dog assessments for predicting behaviour post-adoption [20–24], resulting in calls for longitudinal, observational models of assessment [20,24]. Animal shelters are dynamic environments and, for most dogs, instigate an immediate threat to homeostasis as evidenced by heightened hypothalamic-pituitary-adrenal axis activity and an increase in stress-related behaviours (e.g. [25–28]). Over time, physiological and behavioural responses are amenable to change [17,27,29]. Therefore, dogs in shelters may exhibit substantial heterogeneity in intra-individual behaviour captured neither by standardised behavioural assessments conducted at one time point [24] nor by group-level patterns of behavioural change. An additional complication is that the behaviour in shelters may not be representative of behaviour outside of shelters. For example, Patronek and Bradley [29] suggested that up to 50% of instances of aggression expressed while at a shelter are likely to be false positives. Such false positives may be captured in estimates of predictability, with individuals departing more from their representative behaviour having higher residual intra-individual variability (lower predictability) than others. Overall, absolute values of behaviour, such as mean trait values across time (i.e. personality), may account for just part of the important behavioural variation needed to understand and predict shelter dog behaviour. While observational models of assessment have been encouraged, methods to systematically analyse longitudinal data collected at shelters into meaningful formats are lacking.

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

The analysis of behavioural reaction norms is dependent on the use of hierarchical statistical models for partitioning variance among individuals [3,5,6]. Given that ordinal data are common in behavioural research, here, we illustrate how similar hierarchical models can be applied to ordinal data using a Bayesian framework (see also [31]). Apart from distinguishing inter- from intra-individual variation, we place particular emphasis on two desirable properties of the hierarchical modelling approach taken here. First, the property of *hierarchical shrinkage* [32] offers an efficacious way of making inferences about individual-level behaviour when data are highly unbalanced and potentially unrepresentative of a dog's typical behaviour. When data are sparse for certain individuals, hierarchical shrinkage means that an individual's parameter estimates (e.g. intercepts) are more similar to, or shrunken, towards the group-level estimates. Secondly, since any prediction of future (dog) behaviour will entail uncertainty, a Bayesian approach is attractive because we can directly obtain a probability distribution of parameter values consistent with the data (i.e. the posterior distribution) for all parameters [32,33]. By contrast, frequentist confidence intervals are not posterior probability distributions and, thus, their interpretation is more challenging when a goal is to understand uncertainty in parameter estimates [32].

Material & Methods

Subjects

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

Shelter environment

Details of the shelter environment have been presented elsewhere [35]. Briefly, the shelter was composed of three different rehoming centres (Table 1): one large inner-city centre based in London (approximate capacity: 150-200 dogs), a medium-sized suburban/rural centre based in Old Windsor (approximate capacity: 100-150 dogs), and a smaller rural centre in Brands Hatch (approximate capacity: 50 dogs). Dogs considered suitable for adoption were housed in indoor kennels (typically about 4m x 2m, with a shelf and bedding alcove; see also [36]). Most dogs were housed individually, and given daily access to an indoor run behind their kennel. Feeding, exercising and kennel cleaning were performed by a relatively stable group of staff members. Dogs received water ad libitum and two meals daily according to veterinary recommendations. Sensory variety was introduced daily (e.g. toys, essential oils, classical music, access to quiet ‘chill-out’ rooms). Regular work hours were from 0800 h to 1700 h each day, with public visitation from 1000 h to 1600 h. Dogs were socialised with staff and/or volunteers daily.

Data collection

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

Behavioural observations in the ‘Interactions with unfamiliar people’ context were recorded using a 13-code ethogram (Table 2). Each behavioural code was subjectively labelled and generally defined, providing a balance between behavioural rating and behavioural coding methodologies. The ethogram represented a scale of behavioural problem severity and assumed adoptability (higher codes indicating higher severity of problematic behaviour/lower sociability), reflected by grouping the 13 codes further into green, amber and red codes (Table 2). Green behaviours posed no problems for adoption, amber behaviours suggested dogs may require some training to facilitate successful adoption but did not pose a danger to people or other dogs, and red behaviours suggested dogs needed training or behavioural modification to facilitate successful adoption and could pose a risk to people or other dogs. A dog's suitability for adoption was, however, based on multiple behavioural observations over a number of days. When registering an observation, the employee selected the highest code in the ethogram that was observed on that occasion (i.e. the most severe level of problematic behaviour was given priority). There were periods when a dog could receive no entries for the context for several days but other times when multiple observations were recorded on the same day, usually when a previous observation was followed by a more serious behavioural event. In these instances, and in keeping with the shelter protocol, we retained the highest (i.e. most severe) behavioural code registered for the context that day. When the behaviours were the same, only one record was retained for that day. This resulted in an average of 5.9 (SD = 3.7; range = 1 to 22) records per dog on responses during interactions with unfamiliar people while at the shelter. For dogs with more than one record, the average number of days between records was 2.8 (SD = 2.2; range = 1 to 29).

Validity & inter-rater reliability

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

Statistical analyses

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

Validity & inter-rater reliability

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

Hierarchical Bayesian ordinal probit model

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

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

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

Hierarchical structure

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

$$y_{ij}^* = \beta_0 + \nu_{0j} + \sum_{p=1}^{P} \beta_{p0} x_{pj} + \left(\beta_1 + \nu_{1j} + \sum_{p=1}^{P} \beta_{p1} x_{pj} \right) day_{ij} + \left(\beta_2 + \nu_{2j} + \sum_{p=1}^{P} \beta_{p2} x_{pj} \right) day_{ij}^2 \qquad (2)$$

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

The SD model was:

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

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

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

$$\mathbf{\Sigma_v} = \begin{bmatrix} \tau_{v_0}^2 & \rho_{v_{01}}\tau_{v_0}\tau_{v_1} & \rho_{v_{02}}\tau_{v_0}\tau_{v_2} & \rho_{v_{03}}\tau_{v_0}\tau_{v_3} \\ \dots & \tau_{v_1}^2 & \rho_{v_{12}}\tau_{v_1}\tau_{v_2} & \rho_{v_{13}}\tau_{v_1}\tau_{v_3} \\ \dots & \dots & \tau_{v_2}^2 & \rho_{v_{23}}\tau_{v_2}\tau_{v_3} \\ \dots & \dots & \dots & \tau_{v_3}^2 \end{bmatrix} \tag{4}$$

The diagonal elements are the variances of the dog-level intercepts, linear slopes, quadratic slopes and residual SDs, respectively, while the covariances fill the off-diagonal elements (only the upper triangle shown), where ρ is the correlation coefficient. In the results, we report τ_{v3} (the SD of dog-level residual SDs) on the original scale, rather than the log-transformed scale, using $\sqrt{e^{2\delta+\tau^2_{v3}}e^{\tau^2_{v3}}-1}$. Likewise, δ was transformed to the median of the original scale by e^{δ}.

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

$$ICC_i = \frac{\tau^2_{v_0} + 2Cov_{v_0,v_1}Day_i + \tau^2_{v_1}Day_i^2 + 2Cov_{v_0,v_2}Day_i^2 + \tau^2_{v_2}Day_i^4 + 2Cov_{v_1,v_2}Day_i^3}{numerator + e^{\delta}} \qquad (5)$$

Equation 5 is an extension of the intra-class correlation calculated from mixed-effect models with a random intercept only [43] to include the variance parameters for, and covariances between, the linear and quadratic effects of day, which were evaluated at specific days of interest. We calculated the ICC for values of -1, 0 and 1 on the standardised day scale, corresponding to approximately the arrival day (day 0), day 8, and day 15. This provided a representative spread of days for most of the dogs in the sample, since there were fewer data available for later days which could lead to inflation of inter-individual differences.

To inspect the degree of rank-order change in sociability across dogs from arrival day compared to specific later days (i.e. whether dogs that were, on average, least sociable on arrival also tended to be least sociable later on), we calculated the 'cross-environmental' correlations [44] between the same days as the ICC. The cross-environmental covariance matrix, $\mathbf{\Omega}$, between the three focal days was calculated as:

$$\mathbf{\Omega} = \mathbf{\Psi K \Psi}' \qquad (6)$$

In equation 6, **K** represents the variance-covariance matrix of the dog-level intercepts and (linear and quadratic) slopes, and **Ψ** is a three-by-three matrix with a column vector of 1s, a column vector containing -1, 0, and 1 defining the day values for the cross-environmental correlations for the linear component, and a column vector containing 1, 0, and 1 defining the day values for the cross-environmental correlations for the quadratic component. Once defined, **Ω** was scaled to a correlation matrix. Finally, to summarise the degree of individual differences in predictability, we calculated the 'coefficient of variation for predictability' as $\sqrt{e^{\tau_{v_3}^2} - 1}$ following Cleasby *et al.* [5].

Prior distributions

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

Model selection & computation

We compared the full model explained above to five simpler models. Starting with the full model, the alternative models included: i) parameters quantifying personality and quadratic and linear plasticity only; ii) parameters quantifying personality and linear plasticity only, with a fixed quadratic effect of day since arrival; iii) parameters quantifying personality only, with fixed linear and quadratic effects of day since arrival; iv) parameters quantifying personality only, with a fixed linear effect of day since arrival; and v) a generalised linear regression with no dog-varying parameters and a linear fixed effect for day since arrival (Figure 1). Models were compared by calculating the widely applicable information criterion (WAIC; [47]) following McElreath [33] (see the R script file). The WAIC is a fully Bayesian information criterion that indicates a model's *out-of-sample* predictive accuracy relative to other plausible models while accounting for model complexity, and is preferable to the deviance information criterion (DIC) because WAIC does not assume multivariate normality in the posterior distribution and returns a probability distribution rather than a point estimate [33]. Thus, WAIC guards against both under- and over-fitting to the data (unlike measures of purely in-sample fit, e.g. R^2).

Models were computed using the probabilistic programming language Stan [45] using the *RStan* package [48] version 2.15.1, which employs Markov chain Monte Carlo estimation using Hamiltonian Monte Carlo (see the R script file and Stan code for full details). We ran four chains of 5,000 iterations each, discarding the first 2,500 iterations of each chain as warm-up, and setting thinning to 1. Convergence was assessed visually using trace plots to ensure chains were well mixed, numerically using the Gelman-Rubin statistic (values close to 1 and < 1.05 indicating convergence) and by inspecting the effective sample size of each parameter. We also used graphical posterior predictive checks to assess model predictions against the raw data, including 'counterfactual' predictions [33] to inspect how dogs would be predicted to behave across the first month of being in the shelter regardless of their actual number of observations or length of stay at the shelter.

To summarise parameter values, we calculated mean (denoted β) and 95% highest density intervals (HDIs), the 95% most probable values for each parameter (using functions in the *rethinking* package; [33]). For comparing levels of categorical variables, the 95% HDI of their differences were calculated (i.e. the differences between the coefficients at each step in the MCMC chain, denoted β_{diff}). When the 95% HDI of predictor variables surpassed zero, a credible effect was inferred.

Results

Inter-rater reliability & validity

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

Hierarchical ordinal probit model

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

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

At the individual level, heterogeneity existed in behavioural trajectories across days since arrival (Figure 2b). The SDs of dog-varying parameters were: i) intercepts: 1.29 (95% HDI: 1.18, 1.41; Figure 3a), ii) linear slopes: 0.56 (95% HDI: 0.47, 0.65; Figure 3b), iii) quadratic slopes: 0.28 (95% HDI: 0.20, 0.35; Figure 3c), and iv) residual SDs: 1.39 (95% HDI: 1.22, 1.58; Figure 3d). There was also large uncertainty in individual-level estimates. Figure 4 displays counterfactual model predictions for twenty randomly-sampled dogs. Uncertainty in reaction norm estimates, illustrated by the width of the 95% HDIs (dashed black lines), was greatest when data were sparse (e.g. towards the end of the one-month study period). Hierarchical shrinkage meant that individuals with observations of less sociable responses, or individuals with few behavioural observations, tended to have model predictions pulled towards the overall mean. Note that regression lines depict values on the latent scale predicted to generate observations on the ordinal scale, and so may not clearly fit the ordinal data points. The coefficient of variation for predictability was 0.64 (95% HDI: 0.58, 0.70). Individuals with the five highest and lowest residual SD estimates are shown in Figure 5.

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

The ICC by day increased from arrival day (ICC = 0.22; 95% HDI: 0.16, 0.28) to day 8 (ICC = 0.33; 95% HDI: 0.28, 0.38) but changed little by day 15 (ICC = 0.32; 95% HDI: 0.27, 0.37). The cross-environmental correlation between days 0 and 8 was 0.79 (95% HDI: 0.70, 0.88), between days 0 and 15 was 0.51 (95% HDI: 0.35, 0.68), and between days 8 and 15 was 0.95 (95% HDI: 0.93, 0.97).

A one SD increase in the number of observations was associated with higher intercepts (β = 0.12; 95% HDI: 0.03, 0.21; see Supplementary Material Table S2) and higher residual SDs (β = 0.06, 95% HDI: 0.02, 0.10). Increasing age by one SD was associated with lower intercepts (β= -0.61, 95% HDI: -0.70, -0.51), steeper linear slopes (β = -0.20, 95% HDI: -0.27, -0.13), a stronger quadratic curve (β= 0.07, 95% HDI: 0.03, 0.12), and larger residual SDs (β = 0.05, 95% HDI: 0.01, 0.09). Increasing weight by one SD was associated with shallower quadratic curves (β = -0.05, 95% HDI: -0.09, -0.01). No credible effect of sex was observed on personality, plasticity or predictability. Gift dogs had larger intercepts than returned dogs (β_{diff} = 0.28, 95% HDI: 0.04, 0.52) and stray dogs (β_{diff} = 0.33, 95% HDI: 0.15, 0.50), as well as steeper linear slopes (β_{diff} = -0.25, 95% HDI: -0.38, -0.13) and higher residual SDs than stray dogs (β_{diff} = 0.10, 95% HDI: 0.02, 0.18). Dogs at the large rehoming centre had steeper linear slopes (β_{diff} = -0.70, 95% HDI: -0.84, -0.56) and stronger quadratic curves (β_{diff} = 0.35, 95% HDI: 0.26, 0.45) than dogs at the medium rehoming centre, and lower intercept parameters (β_{diff} = -0.30, 95% HDI: -0.50, -0.09) and steeper linear slopes (β_{diff} = -0.22, 95% HDI: -0.38, -0.06) than dogs at the small rehoming centre. Compared to dogs at the small rehoming centre, dogs at the medium centre had lower intercepts (β_{diff}= -0.25, 95% HDI: -0.48, -0.01), and shallower linear (β_{diff} = 0.48, 95% HDI: 0.30, 0.66) and quadratic slopes (β_{diff} = -0.34, 95% HDI: -0.46, -0.22). Dogs already neutered before arrival to the shelter had lower intercepts (β_{diff} = -0.54, 95% HDI: -1.07, -0.03) and lower residual SDs (β_{diff} = -0.53, 95% HDI: -0.85, -0.22) than dogs not neutered, but higher intercepts (β_{diff} = 0.20, 95% HDI: 0.03, 0.37) and higher residual SDs (β_{diff} = 0.10, 95% HDI: 0.02, 0.19) than those neutered whilst at the shelter. Unneutered dogs had higher intercepts (β_{diff} = 0.74, 95% HDI: 0.20, 1.26) and higher residual SDs (β_{diff} = 0.63, 95% HDI: 0.30, 0.92) than dogs neutered at the shelter.

Discussion

This study applied the framework of behavioural reaction norms to quantify inter- and intra-individual differences in shelter dog behaviour during interactions with unfamiliar people. This is the first study to systematically analyse behavioural data from a longitudinal, observational assessment of shelter dogs. Dogs demonstrated substantial individual differences in personality, plasticity and predictability, which were not well described by simply investigating how dogs behaved on average. In particular, accounting for individual differences in predictability, or the short-term, day-to-day fluctuations in behaviour, resulted in significant improvement in model fit (Figure 1). Modelling dogs' longitudinal behaviour also demonstrated that behavioural repeatability increased with days since arrival (i.e. increasing proportion of variance explained by between-individual differences), particularly across the first week since arrival. Similarly, while individuals maintained rank-order differences in sociability across smaller periods (i.e. first 8 days), rank-order differences were only moderately maintained between arrival at the shelter and day 15. The results highlight the importance of adopting observational and longitudinal assessments of shelter dog behaviour, provide a method by which to analyse longitudinal data commensurate with other work in animal behaviour, and identify previously unconsidered behavioural measures that could be used to improve the predictive validity of behavioural assessments in dogs.

Average behaviour

At the group-level, dogs' reactions to meeting unfamiliar people were predominantly coded as *Friendly* (Figure 2a), described as 'Dog initiates interactions in an appropriate social manner'. Although this definition is broad, it represents a functional qualitative characterisation of behaviour suitable for the purposes of the shelter when coding behavioural interactions, and its generality may partly explain why it was the most prevalent category. The results are consistent with findings that behaviours indicative of poor welfare and/or difficulty of coping (e.g. aggression) are relatively infrequent even in the shelter environment [22,26]. The change of behaviour across days since arrival was characterised by an increase in the *Friendly* code and a decrease in other behavioural codes (Figure 2a). Furthermore, the positive quadratic effect of day since arrival on sociability illustrates that the rate of behavioural change was not constant across days, being quickest earlier after arrival (Figure 2b). The range of behavioural change at the group-level was, nevertheless, still concentrated around the lowest behavioural codes, *Friendly* and *Excitable*.

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

Individual-level variation

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

Certain studies on dog personality have explored how personality trait scores change across time periods, such as ontogeny (e.g. [53]) or time at a shelter (e.g. [17]). Such analyses assume, however, that individuals have similar degrees of change through time. If individuals differ in the magnitude or direction of change (i.e. degree of plasticity), group-level patterns of change may not capture important individual heterogeneity. In this study, most dogs were likely to show lower behavioural codes/more sociable responses across days since arrival, although the rate of linear and quadratic change differed among dogs. Indeed, some dogs showed a *decrease* in sociability through time (individuals with positive model estimates in Figure 3b), and while most dogs showed greater behavioural change early after arrival, others showed slower behavioural change early after arrival (individuals with negative model estimates in Figure 3c). As with estimates of personality, there was also large uncertainty of plasticity.

Part of the difficulty of estimating reaction norms for heterogeneous data is choosing a function that best describes behavioural change. We examined both linear and quadratic effects of day since arrival based on preliminary plots of the data, and their inclusion in the best fitting full model is supported by the lower WAIC value of alternative model 3, with both effects, compared to 4, with just the linear effect (Figure 1). Most studies are constrained to first-order polynomial reaction norms through time due to collecting data at only a few time points [6,44]. However, the quadratic function was relatively easy to vary across individuals while maintaining interpretability of the results. More complex functions (e.g. regression splines) have the disadvantage of being less easily interpretable and higher-order polynomial functions may produce only crude representations of data-generating processes [33]. Nevertheless, by collecting data more intensely, the opportunities to model behavioural reaction norms beyond simple polynomial effects of time should improve. For instance, ecological momentary assessment studies in psychology point to possibilities for modelling behaviour as a dynamic system, such as with the use of vector-autoregressive models and dynamic network or factor models (e.g. [54,55]). These models can also account for relationships between multiple dependent variables (e.g. multiple measures of sociability). Models of behavioural reaction norms, by contrast, have usually been applied to only one dependent variable operationally defined as reflecting the trait of interest, so methods to model multiple dependent variables through time concurrently will be an important advancement.

Personality and plasticity were correlated, with dogs with less sociable behaviour across days being less plastic. Previous studies have explored the relationship between how individuals behave on average and their degree of behavioural change. David *et al.* [56] found that male golden hamsters (*Mesocricetus auratus*) showing high levels of aggression in a social intruder paradigm were slower in adapting to a delayed-reward paradigm. In practice, the relationship between personality and plasticity is probably context dependent. Betini and Norris [57] found, for instance, that more aggressive male tree swallows (*Tachycineta bicolor*) during nest defence were more plastic in response to variation in temperature, but that plasticity was only advantageous for nonaggressive males and no relationship was present between personality and plasticity in females. The correlation between personality and plasticity indicates a ‘fanning out’ shape of the reaction norms through time (Figure 2b). Consequently, behavioural repeatability or the amount of variance explained by between-individual differences increased as a function of day, but only after the first week after arrival. The 'cross-environmental' correlation, moreover, indicated that the most sociable dogs on arrival day were not necessarily the most sociable on later days at the shelter. In particular, the correlation between sociability scores on arrival day and day 15 was only moderate, supporting Brommer [44] that the rank-ordering of trait scores is not always reliable. By contrast, the cross-environmental correlations between days 0 and 8, and between days 8 and 15, were much stronger. These results suggest that shelters using standardised behavioural assessments would benefit from administering such tests as late as possible after dogs arrive.

Of particular interest was predictability or the variation in dogs' residual SDs. Studies of dog personality generally treat behaviour as probabilistic, implying recognition that residual intra-individual behaviour is not completely stable, and authors have posited that dogs may vary in their behavioural consistency (e.g. [13]). Yet, this is the first study to quantify individual differences in predictability in dogs. Modelling residual SDs for each dog resulted in a model with markedly better out-of-sample predictive accuracy (Figure 1). The coefficient of variation for predictability was 0.64 (95% HDI: 0.58, 0.70), which is high compared to other studies in animal behaviour. For instance, Mitchell *et al.* [6] reported a value of 0.43 (95% HDI: 0.36, 0.53) in spontaneous activity measurements of male guppies (*Poecilia reticulata*). Variation in predictability also supports the hypothesis that dogs have varying levels of behavioural consistency. It is important to note, however, that interactions with unfamiliar people at the shelter were likely more heterogeneous than behavioural measures from standardised tests or laboratory environments, which may contribute to greater individual variation in predictability. Moreover, the behavioural data analysed here may have contained more measurement error than data from more standardised environments.

Although shelter employees demonstrated significant inter-rater reliability in video coding sessions, the average proportion of shelter employees who selected the correct behavioural code to describe behaviours seen in videos was modest (66%), while 78% chose a video in the correct colour category (green, amber or red). Indeed, only 55% of employees identified the *Reacts to people aggressive* behaviour as a red code, with the remaining employees identifying it as the amber category code *Reacts to people non-aggressive*. As discussed by Goold and Newberry [35], employees were likely to mistake examples of aggression for non-aggression, but not the other way around. In the current study, this would have increased the percentage of lower category codes (describing greater sociability). Due to lower standardisation of the observational contexts at the shelter than in formal behavioural testing, it was important to evaluate the reliability and validity of the behavioural records. Defining acceptable standards of reliability and validity is, however, non-trivial and we could not find measures of reliability or validity in any previous studies investigating predictability in animals for comparison.

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

Predictors of individual variation

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

Conclusion

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

Ethics statement

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

Data accessibility

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

Competing interests

We declare no competing interests.

Author contributions

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

Acknowledgements

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

Funding statement

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

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

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

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

Figure 2. (a) Predicted probabilities (posterior means = black lines; 95% highest density intervals = shaded areas) of different sociability codes across days since arrival. (b) Posterior mean behavioural trajectories on the latent scale (ranging from $\pm\infty$) at the group-level (blue line) and for each individual (black lines), where higher values indicate lower sociability.

Figure 3. Posterior means (black dots) and 95% highest density intervals (grey vertical lines) for each dogs' (a) intercept, (b) linear slope, (c) quadratic slope, and (d) residual SD parameter.

Figure 4. Predicted reaction norms ('counterfactual' plots) for twenty randomly-selected dogs. Black points show raw data on the ordinal scale (higher values indicate lower sociability), and solid and dashed lines illustrate posterior means and 95% highest density intervals. When data were sparse, there was increased uncertainty in model predictions. Due to hierarchical shrinkage, individual dogs' model predictions were pulled towards the group-level mean, particularly for those dogs showing higher behavioural codes (i.e. less sociable responses).

Figure 5. Reaction norms (posterior means = solid black lines; 95% highest density intervals = dashed black lines) for individuals with the five highest (top row) and five lowest (bottom row) residual SDs. Black points represent raw data on the ordinal scale (higher values indicating lower sociability).

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

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

1. Bell, A. M., Hankison, S. J. & Laskowski, K. L. 2009. The repeatability of behaviour: a meta-analysis. *Anim. Behav.* **77**, 771–783. (doi: 10.1016/j.anbehav.2008.12.022.□)

2. Biro, P. A., Adriaenssens, B., Cole, A. E. B. J. & Bronstein, E. J. L. 2013. Predictability as a personality trait: consistent differences in intraindividual behavioral variation. *Am. Nat.* **182**, 621–629. (doi:10.1086/673213)

3. Dingemanse, N. J., Kazem, A. J. N., Réale, D. & Wright, J. 2010. Behavioural reaction norms: animal personality meets individual plasticity. *Trends Ecol. Evol.* **25**, 81–89. (doi:10.1016/j.tree.2009.07.013)

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

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

6. Mitchell, D. J., Fanson, B. G., Beckmann, C. & Biro, P. A. 2016 Towards powerful experimental and statistical approaches to study intraindividual variability in labile traits. *Open Sci.* **3**, 160352. (doi:10.1098/rsos.160352)

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

8. Westneat, D. F., Wright, J. & Dingemanse, N. J. 2015 The biology hidden inside residual within-individual phenotypic variation. *Biol. Rev.* **90**, 729–743. (doi:10.1111/brv.12131)

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

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

11. Tonsor, S. J., Elnaccash, T. W. & Scheiner, S. M. 2013 Developmental instability is genetically correlated with phenotypic plasticity, constraining heritability, and fitness. *Evolution* **67**, 2923–2935. (doi:10.1111/evo.12175)

12. Oates, A. C. 2011 What's all the noise about developmental stochasticity? *Development* **138**, 601–607. (doi:10.1242/dev.059923)

13. Fratkin, J. L., Sinn, D. L., Patall, E. A. & Gosling, S. D. 2013 Personality consistency in dogs: a meta-analysis. *PLOS ONE* **8**, e54907. (doi:10.1371/journal.pone.0054907)

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

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

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

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

18. Serpell, J. A. & Duffy, D. L. 2016 Aspects of juvenile and adolescent environment predict aggression and fear in 12-month-old guide dogs. *Font. Vet. Sci.* **3**. (doi:10.3389/fvets.2016.00049)

19. Robinson, L. M., Thompson, R. S. & Ha, J. C. 2016 Puppy temperament assessments predict breed and American Kennel Club group but not adult temperament. *J. Appl. Anim. Welf. Sci.* **19**, 101–114. (doi:10.1080/10888705.2015.1127765)

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

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

22. Mornement, K. M., Coleman, G. J., Toukhsati, S. R. & Bennett, P. C. 2015 Evaluation of the predictive validity of the Behavioural Assessment for Re-homing K9's (B.A.R.K.) protocol and owner satisfaction with adopted dogs. *Appl. Anim. Behav. Sci.* **167**, 35–42. (doi:10.1016/j.applanim.2015.03.013)

23. Poulsen, A. H., Lisle, A. T. & Phillips, C. J. C. 2010 An evaluation of a behaviour assessment to determine the suitability of shelter dogs for rehoming. *Vet. Med. Int.* **2010**, e523781. (doi:10.4061/2010/523781)

24. Rayment, D. J., Groef, B. D., Peters, R. A. & Marston, L. C. 2015 Applied personality assessment in domestic dogs: limitations and caveats. *Appl. Anim. Behav. Sci.* **163**, 1–18. (doi:10.1016/j.applanim.2014.11.020)

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

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

27. Stephen, J. M. & Ledger, R. A. 2005 An audit of behavioral indicators of poor welfare in kenneled dogs in the United Kingdom. *J. Appl. Anim. Welf. Sci.* **8**, 79–95. (doi:10.1207/s15327604jaws0802_1)

28. Rooney, N. J., Gaines, S. A. & Bradshaw, J. W. S. 2007 Behavioural and glucocorticoid responses of dogs (*Canis familiaris*) to kennelling: investigating mitigation of stress by prior habituation. *Physiol. Behav.* **92**, 847–854. (doi:10.1016/j.physbeh.2007.06.011)

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

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

31. Martin, J. G. A., Pirottay, E., Petellez, M. B. & Blumstein, D. T. 2017 Genetic basis of between-individual and within-individual variance of docility. *J. Evol. Biol.* (doi:10.1111/jeb.13048)

32. Kruschke, J. 2014 *Doing bayesian data analysis: A tutorial with r, jags, and stan.* Academic Press.

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

34. Voith, V. L. et al. 2013 Comparison of visual and DNA breed identification of dogs and inter-observer reliability. *American Journal of Sociological Research* **3**, 17–29. (doi: 10.1080/10888700902956151)

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

36. Owczarczak-Garstecka, S. C. & Burman, O. H. 2016 Can sleep and resting behaviours be used as indicators of welfare in shelter dogs (*Canis lupus familiaris*)? *PLOS ONE* **11**, e0163620. (doi: https://doi.org/10.1371/journal.pone.0163620)

37. R Development Core Team 2016 R: A language and environment for statistical computing. Vienna, Austria.

38. Tastle, W. J. & Wierman, M. J. 2007 Consensus and dissention: a measure of ordinal dispersion. *Int. J. Approx. Reason.* **45**, 531–545. (doi:10.1016/j.ijar.2006.06.024)

39. Ruedin, D. 2016 agrmt: Calculate Agreement or Consensus in Ordered Rating Scales. R package version 1.40.4.

40. Liddell, T. M. & Kruschke, J. K. 2015 Analyzing ordinal data: support for a Bayesian approach. *SSRN.* (doi: http://dx.doi.org/10.2139/ssrn.2692323)

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

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

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

44. Brommer, J. E. 2013 Variation in plasticity of personality traits implies that the ranking of personality measures changes between environmental contexts: calculating the cross-environmental correlation. *Behav. Ecol. Sociobiol.* **67**, 1709–1718.

45. Stan Development Team 2016 Stan modeling language users guide and reference manual. Version 2.15.0.

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

47. Watanabe, S. 2010 Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**,

3571–3594. (url: http://www.jmlr.org/papers/v11/watanabe10a.html)

48. Stan Development Team 2016 Rstan: R Interface to Stan. R package version 2.14.1.

49. Hiby, E. F., Rooney, N. J. & Bradshaw, J. W. S. 2006 Behavioural and physiological responses of dogs entering re-homing kennels. *Physiol. Behav.* **89**, 385–391. (doi:10.1016/j.physbeh.2006.07.012)

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

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

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

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

54. Cramer, A. O. J., Borkulo, C. D. van, Giltay, E. J., Maas, H. L. J. van der, Kendler, K. S., Scheffer, M. & Borsboom, D. 2016 Major depression as a complex dynamic system. *PLOS ONE* **11**, e0167490. (doi:10.1371/journal.pone.0167490)

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

56. David, J. T., Cervantes, M. C., Trosky, K. A., Salinas, J. A. & Delville, Y. 2004 A neural network underlying individual differences in emotion and aggression in male golden hamsters. *Neuroscience* **126**, 567–578. (doi:10.1016/j.neuroscience.2004.04.031)

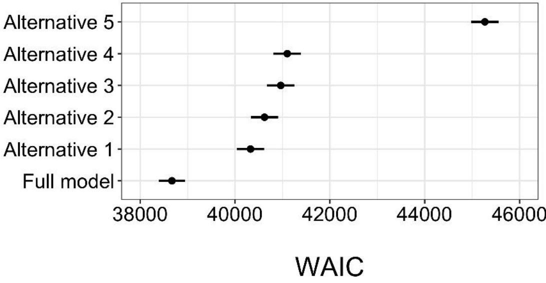
57. Betini, G. S. & Norris, D. R. 2012 The relationship between personality and plasticity in tree swallow aggression and the consequences for reproductive success. *Anim. Behav.* **83**, 137–143. (doi:10.1016/j.anbehav.2011.10.018)

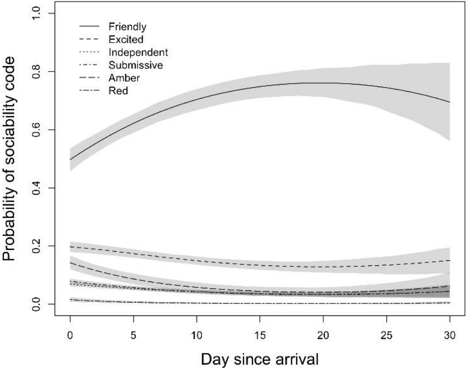
58. Dewitt, T. J., Sih, A. & Wilson, D. S. 1998 Costs and limits of phenotypic plasticity. *Trends Ecol. Evol.* **13**, 77–81.

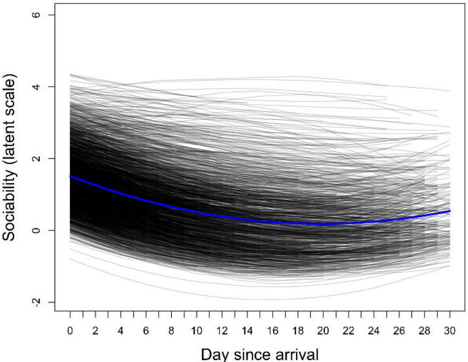
59. Highcock, L. & Carter, A. J. 2014 Intraindividual variability of boldness is repeatable across contexts in a wild lizard. *PLOS ONE* **9**, e95179. (doi:10.1371/journal.pone.0095179)

60. Araya-Ajoy, Y. G. & Dingemanse, N. J. 2017 Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *J. Anim. Ecol.* **86**, 227–238. (doi:10.1111/1365-2656.12621)

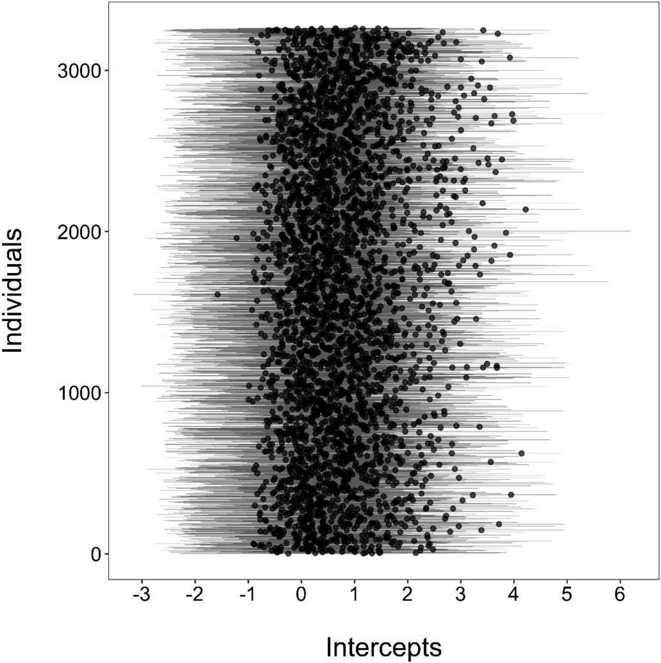
61. Dykiert, D., Der, G., Starr, J. M. & Deary, I. J. 2012 Age differences in intra-individual variability in simple and choice reaction time: systematic review and meta-analysis. *PLOS ONE* **7**, e45759. (doi:10.1371/journal.pone.0045759)

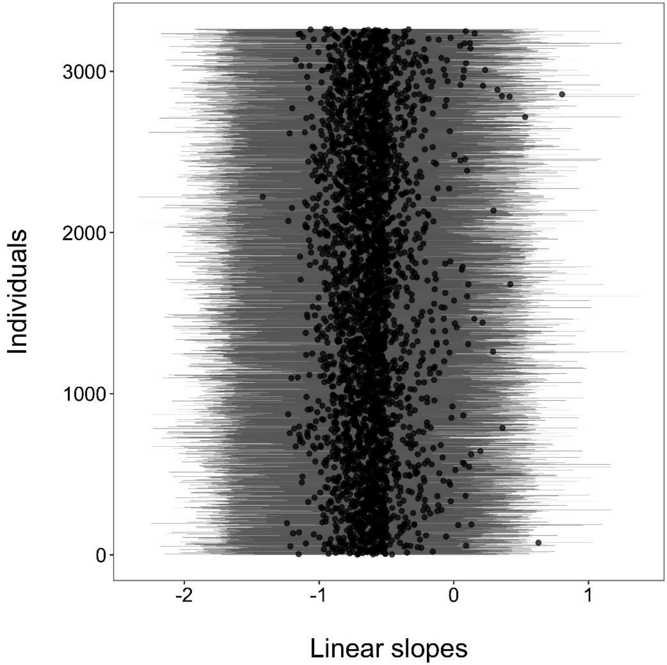

Alternative 5
Alternative 4
Alternative 3
Alternative 2
Alternative 1
Full model
38000
40000
42000
44000
46000
WAIC

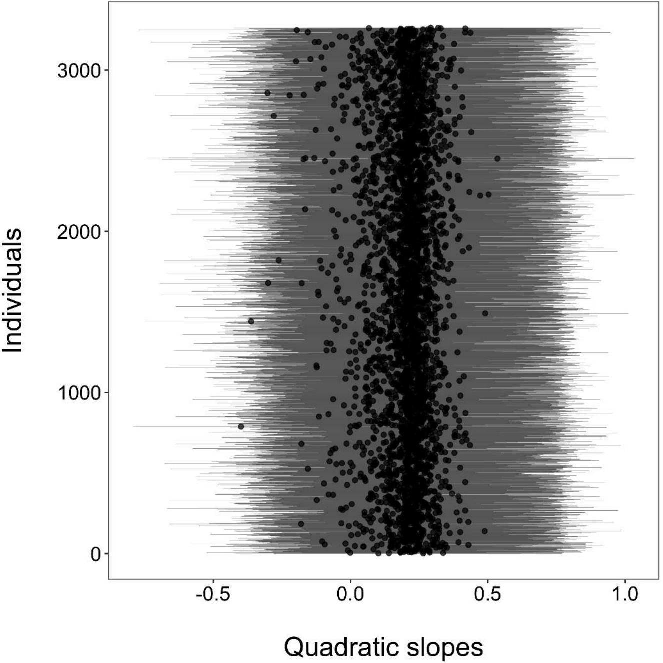

Friendly
Excited
Independent
Submissive
Amber
Red
Probability of sociability code
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Day since arrival



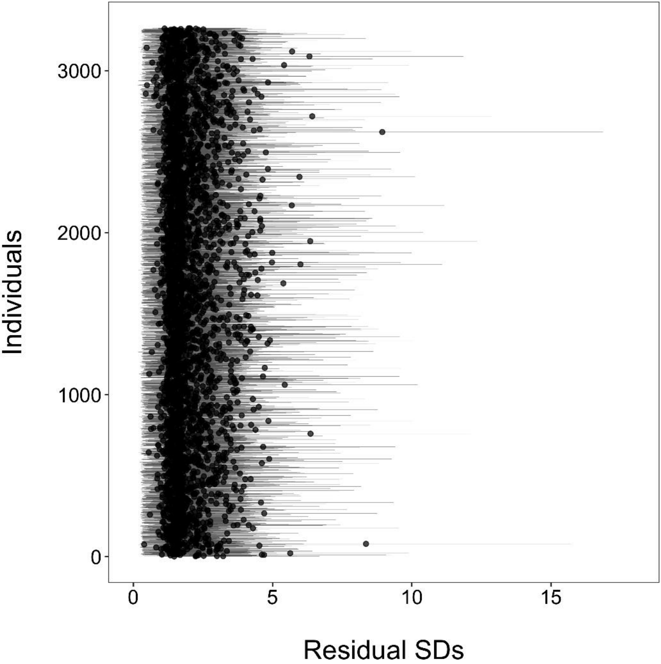
Sociability (latent scale)
6
4
2
0
-2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Day since arrival

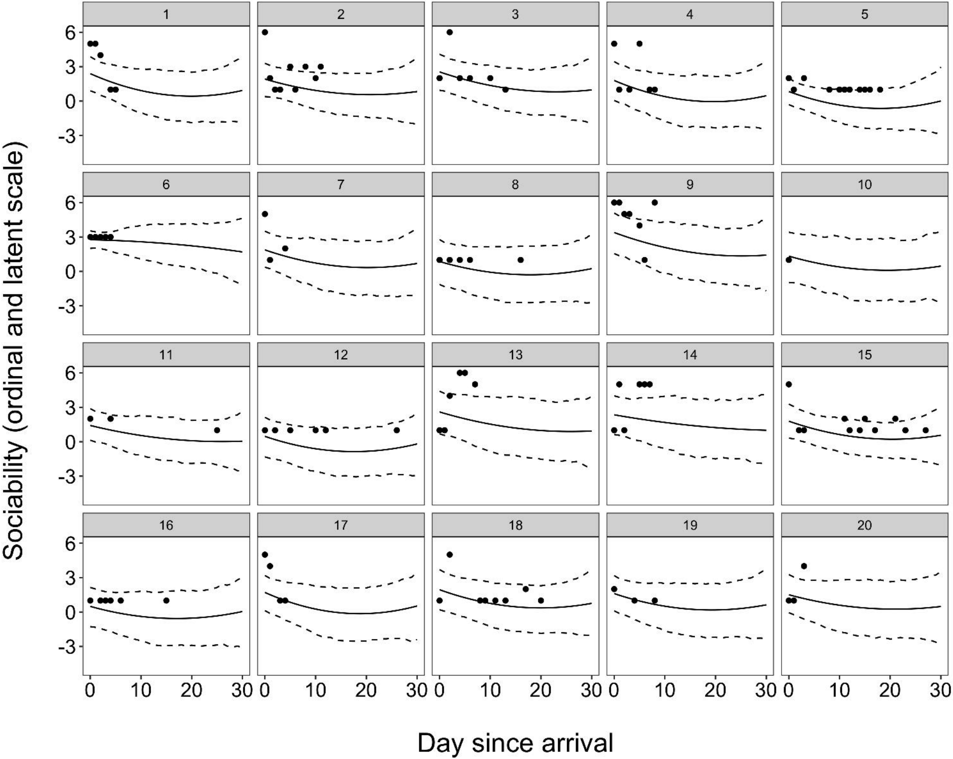

Individuals
0
1000
2000
3000
Intercepts
-3
-2
-1
0
1
2
3
4
5
6


Individuals
3000
2000
1000
0
-2
-1
0
1
Linear slopes

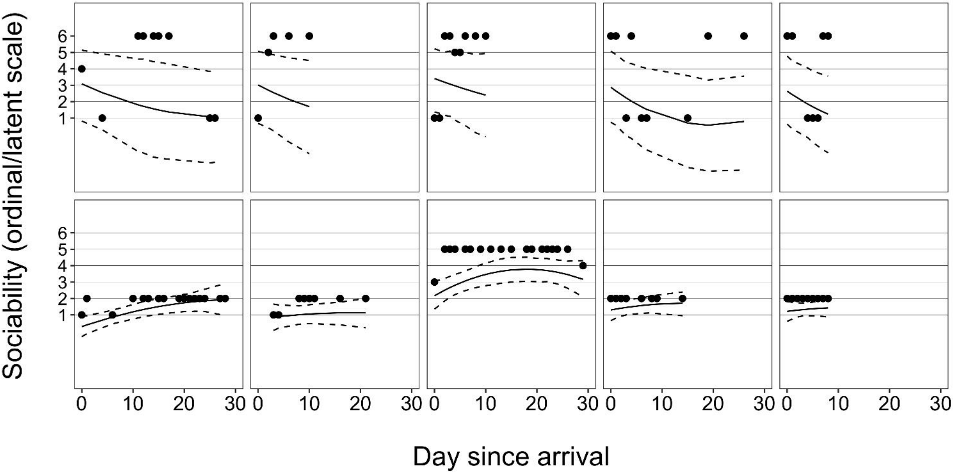


Individuals
3000
2000
1000
0
-0.5
0.0
0.5
1.0
Quadratic slopes


Individuals
3000
2000
1000
0
0
5
10
15
Residual SDs



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
6
3
0
-3
0
10
20
30
Sociability (ordinal and latent scale)
Day since arrival


Sociability (ordinal/latent scale)
6
5
4
3
2
1
Day since arrival
0
10
20
30

Using network analysis to study behavioural phenotypes: an example using domestic dogs

ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Goold C, Vas J, Olsen C, Newberry RC. 2016 Using network analysis to study behavioural phenotypes: an example using domestic dogs. *R. Soc. open sci.* **3**: 160268.
http://dx.doi.org/10.1098/rsos.160268

Received: 18 April 2016
Accepted: 14 September 2016

Subject Category:
Biology (whole organism)

Subject Areas:
behaviour

Keywords:
phenotypic integration, network analysis, dog behaviour, personality, self-organization, play behaviour

Author for correspondence:
Conor Goold
e-mail: conor.goold@nmbu.no

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3500385.

THE ROYAL SOCIETY
PUBLISHING

Using network analysis to study behavioural phenotypes: an example using domestic dogs

Conor Goold[1], Judit Vas[1], Christine Olsen[2] and Ruth C. Newberry[1]

[1]Department of Animal and Aquacultural Sciences, and [2]Section of Public Health, Department of Landscape, Architecture and Spatial Planning, Norwegian University of Life Sciences, Ås, Norway

CG, 0000-0002-9198-0889; JV, 0000-0001-8195-8293;
CO, 0000-0002-9630-1099; RCN, 0000-0002-5238-6959

Phenotypic integration describes the complex interrelationships between organismal traits, traditionally focusing on morphology. Recently, research has sought to represent behavioural phenotypes as composed of quasi-independent latent traits. Concurrently, psychologists have opposed latent variable interpretations of human behaviour, proposing instead a network perspective envisaging interrelationships between behaviours as emerging from causal dependencies. Network analysis could also be applied to understand integrated behavioural phenotypes in animals. Here, we assimilate this cross-disciplinary progression of ideas by demonstrating the use of network analysis on survey data collected on behavioural and motivational characteristics of police patrol and detection dogs (*Canis lupus familiaris*). Networks of conditional independence relationships illustrated a number of functional connections between descriptors, which varied between dog types. The most central descriptors denoted desirable characteristics in both patrol and detection dog networks, with 'Playful' being widely correlated and possessing mediating relationships between descriptors. Bootstrap analyses revealed the stability of network results. We discuss the results in relation to previous research on dog personality, and benefits of using network analysis to study behavioural phenotypes. We conclude that a network perspective offers widespread opportunities for advancing the understanding of phenotypic integration in animal behaviour.

1. Introduction

Understanding the biological organization of complex phenotypes is a mainstay of evolutionary biology [1–3]. Phenotypic integration describes the 'pattern of functional, developmental and/or genetic correlation . . . among different traits in a given organism' [4, p. 266]. Most commonly, phenotypic integration has been concerned with morphological traits (e.g. beak length and size in Darwin's finches [5]; sexual traits [6]). Recently, organization of the behavioural phenotype has also been cast in terms of phenotypic integration. Araya-Ajoy & Dingemanse [7], inspired by research in human psychology, propose that behavioural phenotypes consist of a collection of latent variables (behavioural *characters*) that play a causal role in producing correlated responses in patterns of behaviour, both within and between individuals. They discuss how this conceptualization could be applied to a number of topical themes in the study of animal behaviour including personality (consistent between-individual differences in behaviour) [8] and behavioural plasticity (between-individual differences in behavioural change) [9].

Phenotypic integration of biological traits is increasingly envisaged as interactions (whether physical or correlational) between modules played out on complex networks [10]. For example, Perez *et al.* [11] demonstrate how landmarks of the mammalian mandible can be represented as a network of nodes and correlational edges. Moreover, Wilkins *et al.* [6] advocate a 'phenotype network' approach for understanding correlations between sexual traits in the North American barn swallow (*Hirundo rustica erythrogaster*). The merit of a network perspective is that it naturally incorporates interactions within the components of, and between different, functional traits ('trait complexes') [12] and provides novel analytical insights (e.g. global and local network metrics). This is commensurate with studying organisms as 'developmentally, functionally and phenotypically integrated complex units' [13, p. 279], which numerous researchers argue is integral to improving our knowledge of the phenotype [3,14–16]. It follows that the organization of the behavioural phenotype can also benefit from being represented as an integrated network.

A network perspective has recently emerged in human psychology [17]. Psychological phenomena, such as personality dimensions (e.g. the Five Factor Model) [18], have traditionally been represented as latent variables and analysed with principal components analysis or varieties of factor analysis, respectively. However, this latent variable formulation has been contested (e.g. see commentaries in [19]), based on long-standing concerns that latent variable approaches can be conceptually, statistically and empirically ambiguous [20–24]. The central criticisms are that: (i) latent variables are often represented as fixed entities, failing to portray the dynamics of individual patterns of behaviours and the variability or lack of unidimensionality in psychological variables [23,25,26], (ii) observed behaviours are treated as passive and exchangeable indicators of the particular latent state [27,28], (iii) finding realizations of latent variables in biological organization (e.g. intelligence) [22] is challenging and, more conceptually, (iv) latent variables are unobservable by definition [28,29], promoting circularity in definitions of psychological phenomena ('verbal magic') [21] and leading to the fallacy of misplaced concreteness [30].

The network approach expounded by Cramer *et al.* [17,19] (see also [31–33]) presents personality and psychopathological phenomena as networks of autonomous and causally related cognitive, affective and behavioural components. These components possess conditional independence relationships, such that variation in one component can result in variation in another component conditional on all other measured components [34,35]. Given this assumption, components are more likely to have causal relationships when they possess a functional relationship, and when multiple components form close connections, functional clusters may emerge. For instance, networks of symptoms (e.g. 'loss of energy' and 'weight/appetite change') in long-term patients with major depression disorder were more densely connected (i.e. had greater network connectivity) than those of remitted patients [27]. Van der Mass *et al.* [22] further show how the positive manifold of general intelligence, defined as the observed correlations between cognitive skills related to intelligence, can be explained (and predicted) by direct mutualistic feedback relationships between those cognitive skills. While relationships between network components are influenced by underlying biological mechanisms (e.g. developmental pathways or genetic covariance [36,37]), the network approach aims to understand the behavioural phenotype as its own causal network of self-organizing components, rather than being comprised of passive indicators of 'common cause' latent variables [17].

In this paper, we synthesize the themes introduced above by exploring direct relationships among different behavioural and motivational characteristics in domestic dogs (*Canis lupus familiaris*). Dogs are useful in this respect because it is possible to gather information efficiently about multiple variables in a range of contexts using surveys directed at dog owners, who interact with their dog on a regular basis and are thus qualified to answer questions about their dog's typical behaviour. Such surveys have been shown to be reproducible and corroborate behavioural observations (e.g. clinical reports) [38]. Until now, multivariate data on dog behaviour (e.g. from surveys or direct behavioural assessments) have usually been analysed using latent variable methods to reduce dimensionality and extract latent behavioural traits, or dimensions of dog personality, that explain the correlations between measured variables. This approach has resulted in the identification of a wide number of possible traits [38–40]. Alas, these putative traits often lack strong predictive validity [41–43], a practical concern when recruitment of suitable dogs for specific human uses depends upon reliable predictions. One possible reason is that predictive power is diminished when traits are overestimated as stable, dissociated constructs rather than components of dynamic integrated phenotypes. Further, after conducting a meta-analysis on behavioural consistency across numerous traits, Fratkin *et al.* [40] emphasized that personality dimensions in dogs may still be changeable in adults and sensitive to environmental and social perturbations. Thus, network analysis may be particularly beneficial when applied to the study of dog behaviour because it takes a bottom-up perspective to analysing direct functional relationships between behavioural components rather than decomposing the phenotype into latent variables.

Below, we apply network analysis to survey data collected from police dog handlers on desirable and undesirable behavioural and motivational descriptors of police patrol and detection dogs. Patrol dogs are selected and trained for diverse tasks, such as patrolling areas, controlling crowds, and tracking and detaining suspects, whereas detection dogs search for contraband, commonly drugs and money. Although studies have explored differences between working and non-working dogs on broad behavioural dimensions [44,45], few have compared different types of working dogs. A better understanding of how police dog behaviour is organized is of practical relevance to dog recruitment and training for specialized duties. Rather than focusing on deriving assumed latent traits as a basis for predicting future performance, we elucidate network structures that represent the behavioural phenotypes of patrol and detection dogs. Although our analyses are primarily exploratory, we expected to find some differences between patrol and detection dog networks due to differences in working duties. To our knowledge, this is the first application of network analysis to understand the behavioural phenotypes of animals.

2. Material and methods

2.1. Subjects

This study was carried out in collaboration with members of the Norwegian Police University College in Kongsvinger, Norway who oversee dog selection and training for the Norwegian police force. Professional police dog handlers ($N = 227$) across Norway were invited to complete an online survey in Norwegian investigating the personality and performance of police dogs. Handlers were requested to fill out one survey for each adult dog they had worked with as a handler. A total of 174 surveys were submitted. Three were removed for pertaining to more than one dog. The remaining responses concerned 171 dogs from 117 handlers (mean ± s.d. survey response per person: 1.46 ± 0.65), including 117 patrol dogs (91 German shepherd dogs; 22 Belgian malinois; 1 rottweiler; 1 giant schnauzer; 1 Belgian tervueren; 1 unrecorded breed) and 54 detection dogs (17 labradors; 12 flat coated retrievers; 8 German shepherd dogs; 8 springer spaniels; 2 Belgian malinois; 2 Welsh springer spaniels; 1 German shepherd dog × Belgian shepherd dog; 1 labrador × German pointer; 1 cocker spaniel; 1 Nova Scotia duck-tolling retriever; 1 unrecorded breed). Breed differences were not explored due to the limiting sample sizes. Dogs were mostly entire ($n = 117$) and male ($n = 149$). Responses were received from 79 male and 17 female handlers (21 did not disclose their sex), aged between 28 and 57 (28–37 years: $n = 18$; 38–47 years: $n = 50$; 48–57 years: $n = 28$; undisclosed: $n = 21$). Handlers had between 1 and 30 years of experience as police dog handlers, and on average had 3.75 (s.d. = 4.64) previous dogs (including pet and working dogs).

2.2. Survey development

Survey questions and instructions were constructed in English, translated to Norwegian and back-translated to English to confirm intended meanings. The 'personality section' of the survey included 43 situational and adjective-based descriptors of police dog behavioural characteristics (electronic supplementary material, table S1). The list of descriptors was developed through (i) discussion with members of the Norwegian Police University College to include desirable and undesirable behavioural characteristics of relevance to police dog handlers, (ii) incorporation of characteristics evaluated in standardized assessments of Norwegian police dog behaviour, and (iii) refinement following pilot tests for comprehensibility. Dog handlers rated how well they agreed with the descriptors as portrayals of their dog's typical behaviour, which ranged from 1 = 'Strongly disagree' to 5 = 'Strongly agree', where 3 = 'Neutral'. Participants could also choose 0 = 'Not relevant/I do not know'. All participants were familiar with the terminology used as descriptions of police dog behaviour.

2.3. Data preparation

All data handling and analysis was conducted using R v. 3.2.3 [46] (see the electronic supplementary files for the R script). The raw data for each descriptor contained a mean ± s.d. of 1.23 ± 1.38 (0.72 ± 0.81%) truly missing responses and 4.16 ± 7.64 (2.43 ± 4.47%) zero responses ('Not relevant/I do not know'). Zero responses were particularly prevalent for certain descriptors and five descriptors with at least 10% of zero responses were removed (electronic supplementary material, table S1). The remaining 38 descriptors were all of relevance to both dog types. One handler's responses for a patrol dog were removed as 18.5% were coded as zero (after removal of the five descriptors above), whereas the mean ± s.d. of the percentage of zero responses per dog was 1 ± 2.6%. The remaining zero responses were converted to missing values (as these were not comparable to other responses on the 1–5 scale).

2.4. Multiple imputation

Subsequently, a multiple imputation procedure (using *Amelia*) [47] was used to impute missing scores, rather than applying listwise deletion or mean substitution [48–50]. To ensure its robustness, we investigated any further biases in the data. We first considered whether the pattern of missingness in the data was dependent on dog type (i.e. patrol dogs and detection dogs), or on handlers for those submitting multiple surveys on different dogs (see §1 of the electronic supplementary material for statistical details). There were fewer missing values in the patrol than detection dog responses, and differences in the number of missing values varied between handlers. Thus, we included dog type and numerical handler ID as relevant conditioning variables for the multiple imputation procedure. Secondly, we investigated whether any descriptors had too many missing values to impute. The proportion of missing responses advisable for multiple imputation procedures is variable [51], although 5% or less is commonly considered unproblematic whereas greater than 5% [52] or 10% [53] have been reported to bias results. We chose to remove four descriptors with greater than 5% of missing responses (electronic supplementary material, table S1). Finally, we identified five pairs of variables that were theoretically similar and had high correlations relative to the data as a whole (polychoric correlations $> |0.8|$; see §2 of the electronic supplementary material for details), indicating redundancy. Therefore, we removed one descriptor from each pair (retaining the more specific one where evident, on the presumption that it was answered more reliably; electronic supplementary material, table S1). The resulting 29 descriptors had a mean of 1.2 ± 1.03% missing responses.

Subsequently, 15 multiply imputed datasets were generated. We averaged the datasets and rounded any non-integers to integers to produce a single dataset of ordinal responses. We examined the independence of responses to each question by the 44 handlers who filled out surveys for more than one dog. For eight descriptors, a high ratio of between- to within-handler variation indicated that repeated responses by the same handler lacked independence (see §3 of the electronic supplementary material for methods). Therefore, these eight descriptors were removed (electronic supplementary material, tables S1 and S2). Because the descriptor 'Good at catching a ball' had a particularly low variation ratio (defined as the proportion of responses not the mode) relative to other descriptors (mode = 5; variation ratio = 0.124), it was also removed. The final 20 descriptors used for the network analyses are presented in table 1, along with their modes, variation ratios and abbreviations used in the figures below.

2.5. Network analysis

2.5.1. Network construction

Networks were constructed and analysed using the *qgraph* package [54]. To construct networks that represented conditional independence relationships, we used Gaussian graphical models (GGM; see [55,56] for an overview). GGMs have been applied successfully to understand personality and psychopathology symptomatic networks (e.g. [27,57]). We used GGMs employing L_1 lasso penalties (i.e. least absolute shrinkage and selection operator), where the inverse covariance matrix (i.e. the matrix

Table 1. Descriptors used in the network analysis, including their abbreviations, modes and variation ratios (whole sample statistics shown outside parentheses; patrol and detection dog statistics, respectively, shown within parentheses). Descriptors are placed in alphabetical order (see electronic supplementary material, table S1, for ordering used in the survey).

abbreviation	descriptor name	mode	variation ratio
ACT[a]	active and nimble	5 (5; 5)	0.247 (0.284; 0.167)
ADP[a]	adapts to new situations quickly	5 (5; 5)	0.406 (0.414; 0.389)
CUR[a]	curious	5 (5; 5)	0.229 (0.224; 0.241)
DA[b]	aggressive towards other dogs ('Dog aggressive')[c]	4 (4; 1)	0.735 (0.707; 0.685)
FDA[b]	guards food ('Food aggressive')	1 (1; 1)	0.418 (0.414; 0.426)
FIT[a]	physically fit	5 (5; 5)	0.247 (0.293; 0.148)
FL[a]	fearless	5 (5; 5)	0.482 (0.422; 0.611)
FoH[b]	fear of heights	1 (1; 1)	0.461 (0.457; 0.500)
FSH[a]	able to stay focused during searches	5 (5; 5)	0.324 (0.371; 0.222)
GUS[b]	gives up searches quickly	1 (1; 1)	0.553 (0.586; 0.481)
GWL[b]	strong tendency to growl at strangers	1 (1; 1)	0.476 (0.483; 0.463)
PLA[a]	playful	5 (5; 5)	0.200 (0.207; 0.185)
PS[a]	solves problems on own ('Problem solving')	5 (5; 5)	0.353 (0.345; 0.370)
PSV[a]	persevering	5 (5; 5)	0.265 (0.267; 0.259)
REC[a]	comes when called ('Recalls')	5 (5; 5)	0.424 (0.466; 0.333)
SLP[a]	good at walking on slippery surfaces	5 (5; 5)	0.265 (0.302; 0.185)
SOC[a]	socially attached to you	5 (5; 5)	0.200 (0.224; 0.148)
STR[b]	nervous and tense when startled	1 (1; 1)	0.606 (0.552; 0.722)
TOY[a]	willing to give you a toy	5 (5; 5)	0.424 (0.457; 0.352)
WIL[a]	desires to make you happy ('Willing to please')	5 (5; 5)	0.353 (0.397; 0.259)

[a] Desirable descriptor.

[b] Undesirable descriptor.

[c] Brief descriptions used to form some abbreviations are shown in parentheses.

of partial correlations) was subject to regularization through penalized maximum-likelihood estimation. This resulted in a sparse graph with credibly non-zero partial correlations, with partial correlations near zero being shrunk to zero. Regularization was controlled by a parameter $\lambda \in [0, 1]$ [58]. The optimal value of λ was chosen according to the graph with the lowest Extended Bayesian Information Criterion (EBIC) following Foygel & Drton [59] (see also [56]) and implemented in the 'EBICglasso' function in the *qgraph* package. The EBIC criterion was in turn tuned by a parameter $\gamma \in [0, 1]$ that performs best for positive values of γ [59]. We explored the networks over the entire range of γ (by 0.05 increments) and chose the most conservative value of $\gamma = 0.65$, where values above this resulted in empty graphs for the detection dog network. This method optimized specificity in network estimation (i.e. prioritized the elimination of truly non-existent edges) [60]. Because our data were ordinal, we conducted GGM construction and selection using the matrix of polychoric correlations (see the R script file in the electronic supplementary material), which provided the correlations between ordinal variables assumed to have latent continuous distributions.

2.5.2. Centrality analysis

We explored and compared the structures of patrol and detection dog networks using node-level centrality metrics because nodes that are more central are more important for influencing network structure than peripheral nodes. We chose the metrics *betweenness* and *strength* centrality (defined formally for weighted networks in electronic supplementary material, table S3), where node betweenness represents how many shortest paths (i.e. with minimum distance between two nodes) run through a given node and node strength indicates how strongly each node is connected to other nodes [61,62].

Nodes with high betweenness values acted as mediators between indirectly connected nodes, and nodes with high strength values had stronger correlations with other descriptors.

2.5.3. Network comparison and stability

To compare descriptor centrality between patrol and detection dog networks, 2000 non-parametric bootstrap samples for each network were computed (R package: *bootnet*) [63]. Each bootstrap constructed a network of randomly sampled dogs, with replacement. From these bootstrap samples, we calculated the mean centrality of each descriptor (the overall mean of descriptors' mean betweenness and strength values) and these means were compared with Cliff's delta (δ; R package: '*effsize*') [64], a non-parametric effect size ranging between −1 and +1 (see [27]). To explore network stability, we computed bootstrap samples of the networks 2000 times from networks of 3 to 19 nodes (node-wise bootstrapping), and 2000 times from 25% to 95% (at approximately 8% increments) of the original sample sizes (subject-wise bootstrapping; see [65]). This allowed investigating the rank-order consistency of descriptor centrality values and the correlation between centrality values in the bootstrapped networks with the original networks. Confidence intervals on bootstrapped parameters are not reported due to known biases in their estimation [65].

3. Results

3.1. Descriptive network structures

The patrol dog network (figure 1*a*; see association matrix in the electronic supplementary files) had 55 edges (28.95% of possible edges). 'Curious' had strong positive correlations with 'Playful', 'Problem solving' and 'Fearless'. Additional salient positive correlations appeared between: 'Socially attached to you', 'Recalls' and 'Willing to please'; 'Strong tendency to growl at strangers' and 'Food aggressive'; 'Good at walking on slippery surfaces' and 'Physically fit'; 'Active and nimble' and 'Physically fit'; and 'Fearless' and 'Adapts to new situations quickly'. Negative correlations were evident between: 'Fearless' and 'Nervous and tense when startled'; 'Fear of heights' and 'Good at walking on slippery surfaces'; 'Dog aggressive' and 'Willing to please'; 'Food aggressive' and 'Playful'; and 'Gives up searches quickly' with 'Able to stay focused during searches' and 'Willing to please'.

The detection dog network (figure 1*b*; see association matrix in the electronic supplementary files) had 70 edges (36.84% of possible edges). 'Playful' shared salient positive correlations with 'Curious', 'Persevering', 'Adapts to new situations quickly' and 'Problem solving', and was most negatively correlated with 'Gives up searches quickly'. 'Able to stay focused during searches' shared salient positive correlations with 'Socially attached to you', 'Willing to please', 'Adapts to new situations quickly', 'Willing to give you a toy' and 'Active and nimble'. Strong positive correlations were also evident between: 'Fearless' and 'Curious'; 'Fearless' and 'Problem solving'; 'Good at walking on slippery surfaces' with 'Problem solving' and 'Adapts to new situations quickly'; 'Persevering' and 'Physically fit'; 'Willing to please' and 'Socially attached to you'; 'Gives up searches quickly' and 'Food aggressive'; and 'Strong tendency to growl at strangers' with 'Food aggressive' and 'Dog aggressive'. A strong negative correlation was present between 'Curious' and 'Nervous and tense when startled'.

3.2. Network centrality

Most of the desirable descriptors (table 1) had higher observed centrality values compared with undesirable descriptors (figure 2; see electronic supplementary material, table S4, for raw values). In the patrol dog network, 'Playful' had the highest betweenness centrality and 'Curious' the highest strength centrality, whereas 'Playful' had both the highest betweenness and highest strength centrality values in the detection dog network. Across both networks, 'Active and nimble', 'Curious', 'Physically fit', 'Recalls' and 'Good at walking on slippery surfaces' had higher betweenness and strength values in the patrol dog compared to detection dog network (figure 2). In the detection dog network, 'Dog aggressive', 'Able to stay focused during searches', 'Gives up searches quickly', 'Strong tendency to growl at strangers', 'Problem solving', 'Persevering', 'Nervous and tense when startled' and 'Willing to give you a toy' had higher betweenness and strength values than in the patrol dog network.

Across non-parametric bootstrap samples, only certain descriptor centrality differences had strong effect sizes (figure 3; raw values provided in electronic supplementary material, table S5). Mean centrality differences in 'Curious' ($\delta = 0.452$), 'Good at walking on slippery surfaces' ($\delta = 0.290$) and 'Active and

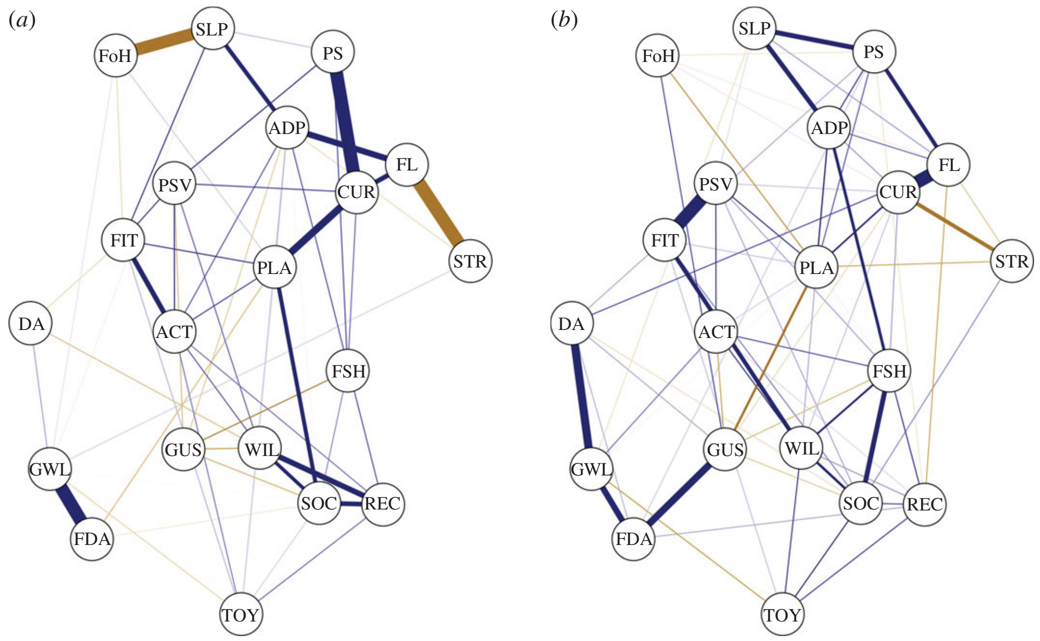


Figure 1. Gaussian graphical models of patrol (*a*) and detection (*b*) dogs. Blue edges show positive correlations, gold edges negative correlations; stronger correlations have thicker edges. See table 1 for descriptor abbreviations.

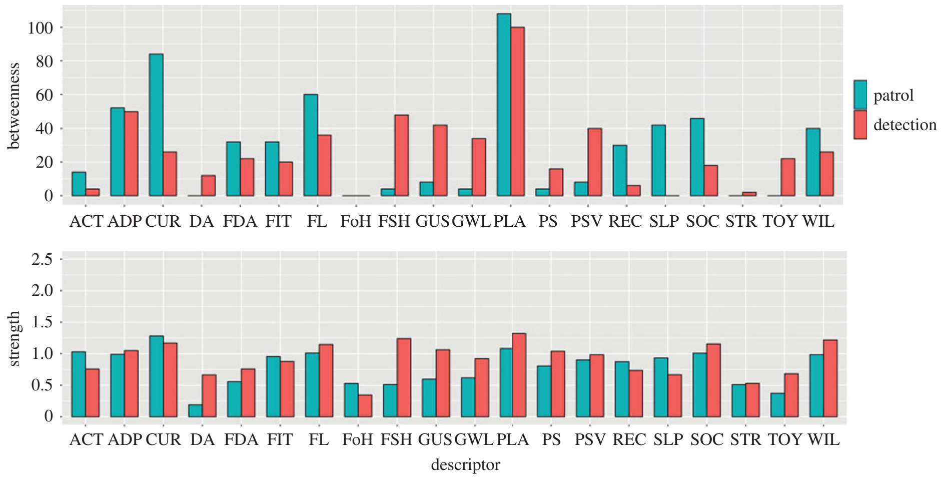


Figure 2. Observed betweenness and strength centrality values (bar heights) for patrol and detection dog networks. See table 1 for descriptor abbreviations and electronic supplementary material, table S4, for raw values.

nimble' ($\delta = 0.282$) had the largest effect sizes in favour of the patrol dog network. 'Able to stay focused during searches' ($\delta = -0.614$), 'Dog aggressive' ($\delta = -0.609$), 'Gives up searches quickly' ($\delta = -0.582$), 'Willing to give you a toy' ($\delta = -0.465$), 'Strong tendency to growl at strangers' ($\delta = -0.310$) and 'Food aggressive' ($\delta = -0.302$) had the largest effect sizes in favour of the detection dog network.

3.3. Network stability

The standard deviation of the number of edges in the patrol dog network across non-parametric bootstrap samples was 12.82, and 29.75 for the detection dog network. Node-wise bootstrapping demonstrated reasonable stability of the original network structures: centrality values from the bootstrapped networks were positively correlated with centrality values in the original networks (figure 4*a*,*b*), even for networks of only three nodes, although the patrol dog network was more stable than the detection dog network (see electronic supplementary material, figure S1, for the rank-order stability of individual descriptors). Network structure was more sensitive under subject-wise bootstrapping. For the patrol dog network, sampled networks of around 60 dogs or less (approximately

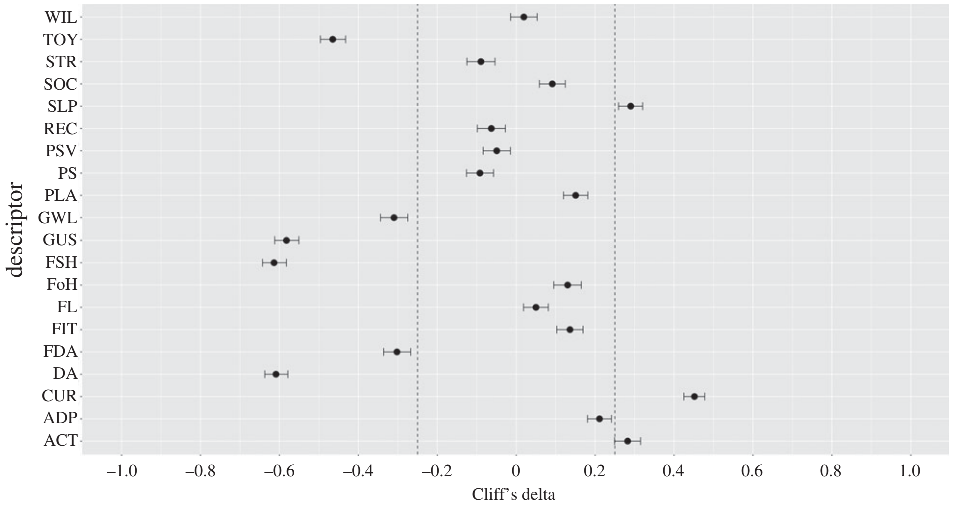


Figure 3. Cliff's delta effect sizes (and 95% CIs) for differences between patrol and detection dog centrality values (average of betweenness and strength) calculated from non-parametric bootstrap samples. Positive values indicate a larger mean for patrol dogs and negative values a larger mean for detection dogs. Values lying within the dashed lines at ±0.25 indicate negligible effect sizes. See table 1 for definitions of descriptor abbreviations and electronic supplementary material, table S5, for raw values.

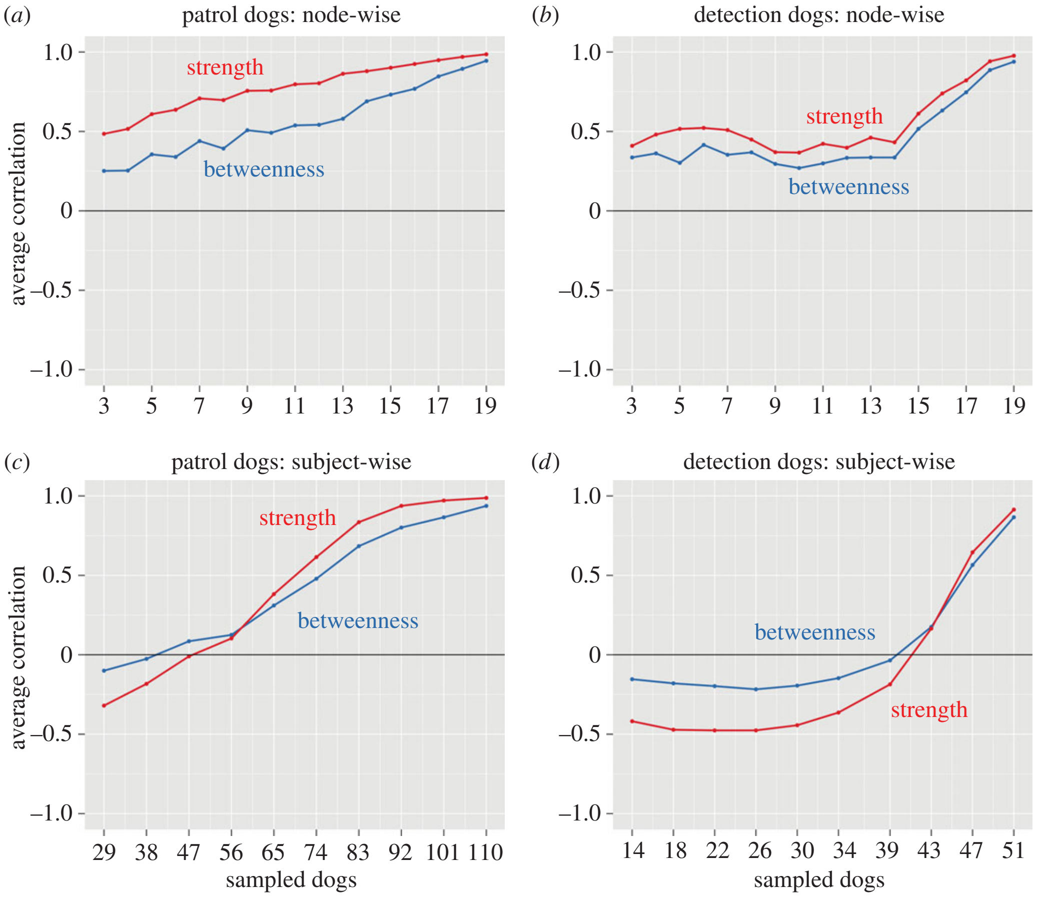


Figure 4. Stability of betweenness and strength centrality values in node-wise (*a*,*b*) and subject-wise (*c*,*d*) bootstrapping. The centrality values in each bootstrapped network were correlated with values in the original networks. Panels (*a*–*d*) show the average correlation across descriptors for each node-wise and subject-wise bootstrap sampling level, respectively.

50% of the original sample size) showed little correlation with the original network values (figure 4*c*). For the detection dog network, networks less than around 40 dogs (approximately 70% of the original sample size) had low to negative correlations with the original network (figure 4*d*; see electronic supplementary material, figure S2, for the rank-order stability of individual descriptors).

4. Discussion

There has been much interest in biology about phenotypic integration of morphological traits, particularly their genetic and developmental bases [2,12,36]. Recent work has extended these notions to conceiving of the behavioural phenotype as composed of quasi-independent latent behavioural traits that form an integrated unit [7]. Network analysis offers benefits for understanding phenotypic integration [6,10,11] and has emerged in human psychology as an efficacious theoretical and analytical framework to understand human behaviour as a causally connected unit [17,32,33,66]. In this regard, it assimilates the study of behavioural phenotypes with research on a number of other complex systems showing how structure can emerge from self-organizing interactions between component parts (e.g. social groups [67], genetic/physiological networks [15,68] and evolutionary processes [14]). In this paper, we have assimilated this cross-disciplinary progression of ideas by using network analysis to understand relationships among behavioural and motivational characteristics in police patrol and detection dogs.

Our analyses revealed numerous direct correlations between functionally related descriptors in both patrol and detection dog networks (figure 1). For instance, behaviours related to aggression (‘Dog aggressive’, ‘Strong tendency to growl at strangers’ and ‘Food aggressive’) were positively correlated, especially in the detection dog network, as were descriptors indicating levels of sociability and/or trainability (‘Socially attached to you’, ‘Willing to please’, ‘Recalls’, ‘Willing to give you a toy’). Moreover, various positive correlations involving ‘Playful’, ‘Curious’, ‘Fearless’ and ‘Socially attached to you’ are partly consistent with Svartberg & Forkman’s [39] interrelated factors ‘playfulness’, ‘curiosity/fearlessness’ and ‘sociability’. Together with ‘chase proneness’, these factors formed a super-trait referred to as ‘boldness’ that was related to working dog performance [44]. Svartberg & Forkman’s [39] results were based on first- and second-order exploratory factor analyses of pairwise correlations, positing boldness as a higher-order latent variable causing covariation between boldness-related behaviours. Our findings extend these results by disentangling potential causal, mutually reinforcing relationships between behaviours. For instance, despite ‘Curious’ and ‘Socially attached to you’ sharing positive pairwise correlations (0.41 and 0.40 for patrol and detection dogs, respectively; see the R script file in the electronic supplementary files for their calculation), they were not directly related in either network of conditional independence relationships, suggesting that their pairwise correlation was due to common mediating variables. For the assessment of dog behaviour, low predictive values of behavioural and personality tests [41–43] may arise from over-estimating the homogeneity of behavioural traits from pairwise correlations when, in fact, trait compositions could be dynamic through time and across contexts. Distinguishing causal relationships from pairwise correlations could refine behavioural assessments through identifying behavioural variables that cause widespread changes in behavioural phenotypes.

Network analysis provides a number of unique metrics to understand patterns of relationships in multivariate data, such as the estimation of network centrality, indicating the relative importance individual components have across network topologies. In particular, the descriptor ‘Playful’ held a central position across networks (figure 2) both in its number of direct behavioural correlations (i.e. strength centrality) with other descriptors, but also in its mediating role between other relationships across the network (i.e. betweenness centrality). Playfulness is postulated to have a positive influence on the success or trainability of working dogs [44,45], comprising part of Svartberg & Forkman’s boldness dimension [39], and has been assayed in working dog assessments by rating a dog’s attentiveness and intensity when engaging in tug-type games with a toy [39,69]. Play also represents a heterogeneous category of behaviour that includes object-related, locomotory and social components [70], and constitutes an important method of reinforcement in training protocols. Thus, from a network viewpoint, playful behaviour may have important causal connections to a wide range of behaviours. In the patrol dog network (figure 1*a*), ‘Playful’ connected additional central descriptors (figure 2), such as between ‘Socially attached to you’ and ‘Curious’ or ‘Fearless’, respectively. In the detection dog network (figure 1*b*), ‘Playful’ had a strong negative relationship with ‘Gives up searches quickly’, the latter being particularly undesirable for detection dogs. As Bradshaw *et al.* [71] review, play in dogs correlates with a number of variables indicating positive well-being, including obedience indicative of close social bonds with owners. Therefore, the centrality of the 'Playful' descriptor in our network analyses holds an interesting organizational position in the behavioural phenotype of police dogs. This organizational role could be further examined in a network framework by quantifying how different forms of playful behaviour relate to other behaviours through time, or between breeds or types of dogs differing systematically in playfulness (e.g. working and pet dogs) [45].

Other descriptors differed in relative centralities between patrol and detection dog networks. In particular, 'Curious' had larger betweenness and strength centrality values in the patrol dog compared with the detection dog network (figure 2), which was also borne out in the non-parametric bootstrap analyses (figure 3). Moreover, 'Good at walking on slippery surfaces' and 'Active and nimble' had larger mean centrality values across bootstrap samples in the patrol dog compared to detection dog networks. By contrast, task-specific descriptors such as 'Able to stay focused during searches' and 'Gives up searches quickly' (which was negatively correlated with desirable descriptors such as 'Playful'; figure 1*b*) were more central in the detection dog network than the patrol dog network, as was 'Willing to give you a toy', which may reflect the tendency for detection dogs to be trained to hold objects gently in their mouths and relinquish objects easily. Descriptors related to aggression were more frequently and strongly negatively correlated with desirable descriptors and positively correlated with undesirable descriptors compared to the patrol dog network. At the same time, weak positive correlations appeared between desirable and undesirable descriptors, such as between 'Dog aggressive' and 'Fearless', 'Recalls' and 'Food aggressive' or 'Socially attached to you' and 'Nervous and tense when startled', which were not present in the patrol dog network. These findings may indicate less stringent behavioural selection criteria for detection dogs compared with patrol dogs, conditional on detection dogs being good at searching. Consequently, successful detection dogs may, on average, be more likely to show correlations between undesirable and desirable behaviours than successful patrol dogs, as long as they show good performance during search tasks.

Nonetheless, our results also demonstrate uncertainty in network structures. Across non-parametric bootstrap samples, the detection dog network had a large standard deviation of estimated edges, probably due to the smaller detection dog sample size. Both networks were relatively stable in response to node-wise bootstrapping (figure 4*a*,*b*; electronic supplementary material, figure S1), but their stability was more sensitive in the subject-wise bootstrapping (figure 4*c*,*d*; electronic supplementary material, figure S2), and so may differ at larger samples sizes. As highlighted by Epskamp *et al.* [65], it is important that network analyses are checked for stability, and that uncertainty in parameter estimates is reported to gauge the predictive accuracy of network models. This is particularly important in dog personality studies employing exploratory analyses of multivariate datasets.

4.1. Limitations and future directions

There are potential limitations to the example presented here. First, the survey descriptors analysed include general behavioural and motivational characteristics (e.g. 'Fearless') that integrate a number of possible behaviour patterns. Thus, this lexical rating approach differs from the quantitative behavioural assays common in, for instance, behavioural ecology research. Nonetheless, rating approaches may be comparable or more beneficial than direct behavioural observations (e.g. in dogs: [72–74]), particularly in cases where raters are highly familiar with the individual animals (see also the discussion in [75]). However, while the survey here was completed by knowledgeable participants and explicated the network approach, no checks of reliability or validity were conducted. Instead, we employed a rigorous data cleaning process, removing 23 of the original 43 descriptors and employing multiple imputation of missing data. Checks of validity have not been fully developed under a network approach [76]. Validity theory attempts to answer whether an indicator measures what it is intended to measure (e.g. whether 'Strong tendency to growl at strangers' measures aggression) and is motivated by a 'reflective' latent variable conceptualization of scientific constructs [77]. However, the network approach does not view indicators, such as the behavioural descriptors analysed here, as measures of latent traits. Instead, the relationship between constructs and indicators is mereological [32,78], such that 'the observables [i.e. indicators] do not measure the construct, but are part of it' [32, p. 5]. Although validity in a network framework is currently in its infancy, exploring how the network approach can refine the predictive validities of current personality tests in dogs would be a fruitful avenue of research.

Secondly, the network analysis reported here was based on one survey per dog. Although handlers responded regarding dogs' typical behaviours, there are advantages to gathering repeated measurements to directly estimate variation between and within individuals. Network analysis can also be applied to this end (e.g. see [79] for a multilevel time-series network model).

Finally, there is a natural relationship between integration of behavioural phenotypes and the study of animal personality and, relatedly, behavioural syndromes. Animal personality is defined by repeatable between-individual differences in behaviour reflecting personality traits [8,80,81]. As in studies of human personality, investigations into animal personality have used latent variable approaches (e.g. exploratory factor analysis [38] or structural equation modelling [72,82] in dogs) to extract relevant traits. However, the conceptualization of personality traits has been a point of confusion in animal behaviour [40,75,83] and psychologists have related a similar ambiguity in human research directly to latent variable interpretations [20,21,23,24]. Combining the network perspective established in human psychology and the more general biological concept of phenotypic integration may improve the clarity of personality definitions. That is, the behavioural phenotype becomes organized through causal connections between its components. By virtue of this organization, consistent behavioural expression is maintained through principles of network stability [84]. In this way, traits are emergent properties of clustering between functionally related behaviours [17,32]. In psychology, dynamic systems approaches to behaviour have a long history [85], supporting the process of behavioural integration as a self-organizing system [86,87]. In evolutionary biology, morphological trait complexes have been elucidated as emergent properties ('evolutionarily stable configurations') [12] and, more recently, Watson *et al.* [14] use principles of supervised and unsupervised learning to outline how phenotypic correlations can become causal connections over evolutionary timescales, highlighting the role of self-organization in the evolution of phenotypic integration.

5. Conclusion

Network analysis provides a novel approach to conceptualizing and analysing the behavioural phenotype, in both humans and animals. Following recent work across the biological study of phenotypic integration and human psychology, network analysis can be used to conceive of the behavioural repertoire of individuals as a connected system of causally dependent components. We have demonstrated how network analysis can be applied using police patrol and detection dogs as an example, elucidating commonalities and differences between networks in the interrelationships between behavioural and motivational descriptors. Moreover, we have demonstrated how analyses can be carried out to ascertain the stability of the results. We conclude that a network approach offers widespread opportunities for advancing the understanding of phenotypic integration in animal behaviour.

Ethics. Ethical approval from the Regional Ethics Committee in Norway was not required for the collection of this data, as it did not involve handling or experimenting on animals. Approval was acquired from the Norwegian Social Science Data services for the handling of personal data (approval no. 44121).

Data accessibility. The electronic supplementary material and the electronic supplementary files (including the complete raw data, the data after multiple imputation, the patrol and detection dog Gaussian graphical model association matrices, and R script file to reproduce the analyses and figures in this article) are available on Dryad Digital Repository at http://dx.doi.org/10.5061/dryad.81k11 [88].

Authors' contributions. J.V., C.O. and R.C.N. collected the data; C.G. conducted the statistical analyses; all authors discussed and interpreted the results; C.G., J.V., and R.C.N wrote the paper. All authors gave final approval for publication.

Competing interests. The authors declare no competing interests.

Funding. All authors were funded through employment at the Norwegian University of Life Sciences.

Acknowledgements. We thank Janne Karin Brodin of IHA, NMBU for help with setting up and running the questionnaire in Questback, Petter Svensen and Jo Geir Grande of the Politihøgskolen Hundetjenesten for facilitating the study, and all the Norwegian police dog handlers who filled out the questionnaire.

References

1. Wagner GP. 1996 Homologues, natural kinds and the evolution of modularity. *Am. Zool.* **36**, 36–43. (doi:10.1093/icb/36.1.36)
2. Wagner GP, Laubichler MD. 2000 Character identification in evolutionary biology: the role of the organism. *Theory Biosci.* **119**, 20–40. (doi:10.1007/s12064-000-0003-7)
3. Murren CJ. 2012 The integrated phenotype. *Integr. Comp. Biol.* **52**, 64–76. (doi:10.1093/icb/ics043)
4. Pigliucci M. 2003 Phenotypic integration: studying the ecology and evolution of complex phenotypes. *Ecol. Lett.* **6**, 265–272. (doi:10.1046/j.1461-0248.2003.00428.x)
5. Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ. 2006 The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* **442**, 563–567. (doi:10.1038/nature04843)
6. Wilkins MR, Shizuka D, Joseph MB, Hubbard JK, Safran RJ. 2015 Multimodal signalling in the North American barn swallow: a phenotype network approach. *Proc. R. Soc. B* **282**, 20151574. (doi:10.1098/rspb.2015.1574)

7. Araya-Ajoy YG, Dingemanse NJ. 2014 Characterizing behavioural 'characters': an evolutionary framework. *Proc. R. Soc. B* **281**, 20132645. (doi:10.1098/rspb.2013.2645)
8. Sih A, Bell AM, Johnson JC, Ziemba RE. 2004 Behavioral syndromes: an integrative overview. *Q. Rev. Biol.* **79**, 241–277. (doi:10.1086/422893)
9. Dingemanse NJ, Wolf M. 2013 Between-individual differences in behavioural plasticity within populations: causes and consequences. *Anim. Behav.* **85**, 1031–1039. (doi:10.1016/j.anbehav.2012.12.032)
10. Wagner GP, Pavlicev M, Cheverud JM. 2007 The road to modularity. *Nat. Rev.* **8**, 921–931. (doi:10.1038/nrg2267)
11. Perez SI, de Aguiar MAM, Guimarães Jr PR, dos Reis SF. 2009 Searching for modular structure in complex phenotypes: inferences from network analysis. *Evol. Biol.* **36**, 416–422. (doi:10.1007/s11692-009-9074-7)
12. Wagner GP, Schwenk K. 2001 Function and the evolution of phenotypic stability: connecting pattern with process. *Am. Zool.* **41**, 552–563.
13. Forsman A. 2015 Rethinking phenotypic plasticity and its consequences for individuals, populations and species. *Heredity* **115**, 276–284. (doi:10.1038/hdy.2014.92)
14. Watson RA *et al.* 2015 Evolutionary connectionism: algorithmic principles underlying the evolution of biological organisation in evo-devo, evo-eco and evolutionary transitions. *Evol. Biol.* 1–29. (doi:10.1007/s11692-015-9358-z)
15. Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER. 2015 Modelling the epigenetic attractors landscape: toward a post-genomic mechanistic understanding of development. *Front. Genet.* **6**, 1–14. (doi:10.3389/fgene.2015.00160)
16. Ellers J, Liefting M. 2015 Extending the integrated phenotype: covariance and correlation in plasticity of behavioural traits. *Curr. Opin. Insect Sci.* **9**, 31–35. (doi:10.1016/j.cois.2015.05.013)
17. Cramer AOJ, van der Sluis S, Noordhof A, Wichers M, Geschwind N, Aggen SH, Kendler KS, Borsboom D. 2012 Dimensions of normal personality as networks in search of equilibrium: you can't like parties if you don't like people. *Euro. J. Pers.* **26**, 414–431. (doi:10.1002/per.1866)
18. McCrae RR, Costa PT. 1995 Trait explanations in personality psychology. *Euro. J. Pers.* **9**, 231–252. (doi:10.1002/per.2410090402)
19. Cramer AOJ, Waldorp LJ, van der Maas HLJ, Borsboom D. 2010 Comorbidity: a network perspective. *Behav. Brain Sci.* **33**, 137–150. (doi:10.1017/S0140525X09991567)
20. Borsboom D. 2005 *Measuring the mind*. Cambridge, UK: Cambridge University Press.
21. Boag S. 2011 Explanation in personality psychology: 'verbal magic' and the five-factor model. *Phil. Psychol.* **24**, 223–243. (doi:10.1080/09515089.2010.548319)
22. van der Maas HLJ, Dolan CV, Grasman RPPP, Wicherts JM, Huizenga HM, Raijmakers EJ. 2006 A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol. Rev.* **113**, 842–861. (doi:10.1037/0033-295X.113.4.842)
23. Cervone D. 2005 Personality architecture: within-person structures and processes. *Annu. Rev. Psychol.* **56**, 423–452. (doi:10.1146/annurev.psych.56.091103.070133)
24. Lamiell JT. 2013 Statisticism in personality psychologists' use of trait constructs: what is it? How was it contracted? Is there a cure? *New Ideas Psychol.* **31**, 65–71. (doi:10.1016/j.newideapsych.2011.02.009)
25. Adolf J, Schuurman NK, Borkenau P, Borsboom D, Dolan CV. 2014 Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. *Front. Psychol.* **5**, 9–22. (doi:10.3389/fpsyg.2014.00883)
26. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. 2016 Measuring depression over time . . . or not? Lack of unidimensionality and longitudinal measurement invariance in four common ratings scales of depression. *Psychol. Assess.* 1–15. (doi:10.1037/pas0000275)
27. Van Borkulo C, Boschloo L, Borsboom D, Phennix BWJH, Waldorp LJ, Schoevers RA. 2015 Association of symptom network structure with the course of longitudinal depression. *JAMA Psych.* **72**, 1219–1226. (doi:10.1001/jamapsychiatry.2015.2079)
28. Bollen KA. 2002 Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* **53**, 605–634. (doi:10.1146/annurev.psych.53.100901.135239)
29. Borsboom D. 2008 Psychometric perspectives on diagnostic systems. *J. Clin. Psychol.* **64**, 1089–1108. (doi:10.1002/jclp.20503)
30. Flynn M. 1997 The concept of intelligence in psychology as a fallacy of misplaced concreteness. *Interchange* **28**, 231–244. (doi:10.1023/A:1007317410814)
31. Borsboom D, Cramer AOJ. 2013 Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* **9**, 91–121. (doi:10.1146/annurev-clinpsy-050212-185608)
32. Schmittmann VD, Cramer AOJ, Waldorp LJ, Epskamp S, Kievet RA, Borsboom D. 2013 Deconstructing the construct: a network perspective on psychological phenomenon. *New Ideas Psychol.* **31**, 43–53. (doi:10.1016/j.newideapsych.2011.02.007)
33. Goekoop R, Goekoop JG, Scholte HS. 2012 The network structure of human personality according to the NEO-PI-R: matching network community structure to factor structure. *PLoS ONE* **7**, e51558 (doi:10.1371/journal.pone.0051558)
34. Costantini G, Epskamp S, Borsboom D, Perugini M, Mõttus R, Waldorp LJ, Cramer AOJ. 2015 State of the aRt personality research: a tutorial on network analysis of personality data. *J. Res. Pers.* **54**, 13–29. (doi:10.1016/j.jrp.2014.07.003)
35. Pearl J. 2009 Causal inference in statistics: an overview. *Stat. Surv.* **3**, 96–146. (doi:10.1214/09-SS057)
36. Gonzalez PN, Hallgrimsson B, Oyheart EE. 2011 Developmental plasticity in covariance structure of the skull: effects of prenatal stress. *J. Anat.* **218**, 243–257. (doi:10.1111/j.1469-7580.2010.01326.x)
37. Santos JC, Cannatella DC. 2011 Phenotypic integration emerges from aposematism and scale in poison frogs. *Proc. Natl Acad. Sci. USA* **108**, 6175–6180. (doi:10.1073/pnas.1010952108)
38. Hsu Y, Serpell JA. 2003 Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *J. Am. Vet. Med. Assoc.* **9**, 1293–1300. (doi:10.2460/javma.2003.223.1293)
39. Svartberg K, Forkman B. 2002 Personality traits in the domestic dog (*Canis familiaris*). *Appl. Anim. Behav. Sci.* **79**, 133–155. (doi:10.1016/S0168-1591(02)00121-1)
40. Fratkin JL, Sinn DL, Patall EA, Gosling SD. 2013 Personality consistency in dogs: a meta-analysis. *PLoS ONE* **8**, e54907. (doi:10.1371/journal.pone.0054907)
41. Mornement KM, Toukhsati S, Coleman GJ, Bennett P. 2015 Evaluation of the predictive validity of the Behavioural Assessment for Re-homing K9's (B.A.R.K.) protocol and owner satisfaction with adopted dogs. *Appl. Anim. Behav. Sci.* **167**, 35–42. (doi:10.1016/j.applanim.2015.03.013)
42. Reimer S, Müller C, Virányi Z, Huber L, Range F. 2014 The predictive value of early behavioural assessments in pet dogs—a longitudinal study from neonates to adults. *PLoS ONE* **9**, e101237. (doi:10.1371/journal.pone.0101237)
43. Sinn DL, Gosling SD, Hillard S. 2010 Personality and performance in military working dogs: reliability and predictive validity of behavioral tests. *Appl. Anim. Behav. Sci.* **127**, 51–65. (doi:10.1016/j.applanim.2010.08.007)
44. Svartberg K. 2002 Shyness–boldness predicts performance in working dogs. *Appl. Anim. Behav. Sci.* **79**, 157–174. (doi:10.1016/S0168-1591(02)00120-X)
45. Asp HE, Fikse WF, Nilsson K, Strandberg E. 2015 Breed differences in everyday behaviour of dogs. *Appl. Anim. Behav. Sci.* **169**, 69–77. (doi:10.1016/j.applanim.2015.04.010)
46. R Core Team. 2015 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org/.
47. Honaker J, King G, Blackwell M. 2011 Amelia II: a program for missing data. *J. Stat. Softw.* **45**, 1–47. (doi:10.18637/jss.v045.i07)
48. Schlomer GL, Bauman S, Card NA. 2010 Best practices for missing data management in counseling psychology. *J. Couns. Psychol.* **57**, 1–10. (doi:10.1037/a0018082)
49. Osborne JW. 2013 *Best practices in data cleaning*. London, UK: Sage Publications, Inc.
50. Rubin DB. 1976 Inference and missing data. *Biometrika* **63**, 581–592. (doi:10.1093/biomet/63.3.581)
51. Dong Y, Peng CJ. 2013 Principled missing data methods for researchers. *SpringerPlus* **2**, 222. (doi:10.1186/2193-1801-2-222)
52. Schafer JL. 1999 Multiple imputation: a primer. *Stat. Methods Med. Res.* **8**, 3–15. (doi:10.1191/096228099671525676)
53. Bennett DA. 2001 How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* **25**, 464–469. (doi:10.1111/j.1467-842X.2001.tb00294.x)
54. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. 2012 qgraph: network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, 1–18. (doi:10.18637/jss.v048.i04)
55. Kolaczyk ER, Csárdi G. 2014 *Statistical analysis of network data with R*. New York, NY: Springer.
56. Yuan M, Lin Y. 2007 Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35. (doi:10.1093/biomet/asm018)
57. Fried EI, Epskamp S, Nesse RM, Tuerlinckx F, Borsboom D. 2016 What are 'good' depression symptoms? Comparing the centrality of DSM and

non-DSM symptoms of depression in a network analysis. *J. Affec. Disord.* **189**, 314–320. (doi:10.1016/j.jad.2015.09.005)

58. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. 2012 The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**, 1059–1069.
59. Foygel R, Drton M. 2010 Extended Bayesian Information Criterion for Gaussian graphical models. *Adv. Neural Inf. Process Syst.* **23**, 2020–2028.
60. Epskamp S. 2016 Regularized Gaussian psychological networks: brief report on the performance of extended BIC model selection. See https://arxiv.org/pdf/1606.05771.
61. Brandes U. 2001 A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**, 163–177. (doi:10.1080/0022250X.2001.9990249)
62. Newman MEJ. 2001 Scientific collaboration II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 1–7. (doi:10.1103/PhysRevE.64.016132)
63. Epskamp S. 2015 bootnet: bootstrap methods for various network estimation routines. R package version 0.3. See https://github.com/SachaEpskamp/bootnet.
64. Torchiano M. 2015 effsize: efficient effect size computation. R package version 0.5.4. See http://www.cran.r-project.org/package=effsize/.
65. Epskamp S, Borsboom D, Fried EI. 2016 Estimating psychological networks and their stability: a tutorial paper. See https://arxiv.org/abs/1604.08462.
66. Robinaugh DJ, LeBlanc NJ, Vuletich HA, McNally RJ. 2014 Network analysis of persistent complex bereavement disorder in conjugally bereaved adults. *J. Abnorm. Psychol.* **123**, 510–522. (doi:10.1037/abn0000002)
67. Sumpter DJT. 2010 *Collective animal behavior*. Princeton, NJ: Princeton University Press.
68. Cohen AA, Martin LB, Wingfield JC, McWilliams SR, Dunne JA. 2012 Physiological regulatory networks: ecological roles and evolutionary constraints. *Trends Ecol. Evol.* **27**, 428–435. (doi:10.1016/j.tree.2012.04.008)
69. Foyer P, Wilsson E, Wright D, Jensen P. 2013 Early experiences modulate stress coping in a population of German shepherd dogs. *Appl. Anim. Behav. Sci.* **146**, 79–87. (doi:10.1016/j.applanim.2013.03.013)
70. Špinka M, Newberry RC, Bekoff M. 2001 Mammalian play: training for the unexpected. *Q. Rev. Biol.* **76**, 141–168. (doi:10.1086/393866)
71. Bradshaw JWS, Pullen AJ, Rooney NJ. 2015 Why do adult dogs ‘play’? *Behav. Process.* **110**, 82–87. (doi:10.1016/j.beproc.2014.09.023)
72. Barnard S, Marshall-Pescini S, Passalacqua C, Beghelli A, Capra A, Normando S, Pelosi A, Valsecchi P. 2016 Does subjective rating reflect behavioural coding? Personality in 2 month-old dog puppies: an open-field test and adjective-based questionnaire. *PLoS ONE* **11**, e0149831. (doi:10.1371/journal.pone.0149831)
73. Wilson E, Sinn DL. 2012 Are there differences between behavioral measurement methods? A comparison of the predictive validity of two rating methods in a working dog program. *Appl. Anim. Behav. Sci.* **141**, 158–172. (doi:10.1016/j.applanim.2012.08.012)
74. Mirkó E, Dóka A, Miklósi Á. 2013 Association between subjective rating and behaviour coding and the role of experience in making video assessments on the personality of the domestic dog (*Canis familiaris*). *Appl. Anim. Behav. Sci.* **149**, 45–54. (doi:10.1016/j.applanim.2013.10.003)
75. Nettle D, Penke L. 2010 Personality: bridging the literatures from human psychology and behavioural ecology. *Phil. Trans R. Soc. B* **365**, 4043–4050. (doi:10.1098/rstb.2010.0061)
76. Cramer AOJ. 2012 Why the item ‘23 + 1’ is not in a depression questionnaire: validity from a network perspective. *Measurement Interdiscip. Res. Perspect.* **10**, 50–54. (doi:10.1080/15366367.2012.681973)
77. Borsboom D, Mellenbergh GJ, van Heerden J. 2004 The concept of validity. *Am. Psychol. Rev.* **111**, 1061–1071. (doi:10.1037/0033-295X.111.4.1061)
78. Borsboom D, Cramer AOJ, Kievet RA, Scholten AZ, Franic S. 2009 The end of construct validity. In *The concept of validity* (ed. RW Lissitz), pp. 135–170. Charlotte, NC: Information Age Publishing.
79. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, Borsboom D, Tuerlinckx F. 2013 A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS ONE* **8**, e60188. (doi:10.1371/journal.pone.0060188)
80. Carter AJ, Feeney WE, Marshall HH, Cowlishaw G, Heinsohn R. 2013 Animal personality: what are behavioral ecologists measuring? *Biol. Rev.* **88**, 465–475. (doi:10.1111/brv.12007)
81. Koski SE. 2014 Broader horizons for animal personality research. *Front. Ecol. Evol.* **2**, 1–6. (doi:10.3389/fevo.2014.00070)
82. Ley JM, Bennett PC, Coleman GJ. 2009 A refinement and validation of the Monash Canine Personality Questionnaire (MCPQ). *Appl. Anim. Behav. Sci.* **116**, 220–227. (doi:10.1016/j.applanim.2008.09.009)
83. David M, Dall SRX. 2016 Unravelling the philosophies underlying ‘animal personality’ studies: a brief re-appraisal of the field. *Ethology* **122**, 1–9. (doi:10.1111/eth.12445)
84. Kauffman SA. 1993 *The origins of order: self-organization and selection in evolution*. Oxford, UK: Oxford University Press.
85. Thelen E, Smith LB. 1994 *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: Massachusetts Institute of Technology Press.
86. Smith LB, Thelen E. 2003 Development as a dynamic system. *Trends Cogn. Sci.* **7**, 343–348. (doi:10.1016/S1364-6613(03)00156-6)
87. Barrett L. 2011 *Beyond the brain: how body and environment shape animal and human minds*. Princeton, NJ: Princeton University Press.
88. Goold C, Vas C, Olsen C, Newberry RC. 2016 Data from: Using network analysis to study behavioural phenotypes: an example using domestic dogs. Dryad Digital Repository. (doi:10.5061/dryad.81k11)