Personality Consistency in Dogs: A Meta-Analysis

Jamie L. Fratkin   , David L. Sinn, Erika A. Patall, Samuel D. Gosling

Abstract

Personality, or consistent individual differences in behavior, is well established in studies of dogs. Such consistency implies predictability of behavior, but some recent research suggests that predictability cannot be assumed. In addition, anecdotally, many dog experts believe that 'puppy tests' measuring behavior during the first year of a dog's life are not accurate indicators of subsequent adult behavior. Personality consistency in dogs is an important aspect of human-dog relationships (e.g., when selecting dogs suitable for substance-detection work or placement in a family). Here we perform the first comprehensive meta-analysis of studies reporting estimates of temporal consistency of dog personality. A thorough literature search identified 31 studies suitable for inclusion in our meta-analysis. Overall, we found evidence to suggest substantial consistency ($r=0.43$). Furthermore, personality consistency was higher in older dogs, when behavioral assessment intervals were shorter, and when the measurement tool was exactly the same in both assessments. In puppies, aggression and submissiveness were the most consistent dimensions, while responsiveness to training, fearfulness, and sociability were the least consistent dimensions. In adult dogs, there were no dimension-based differences in consistency. There was no difference in personality consistency in dogs tested first as puppies and later as adults (e.g., 'puppy tests') versus dogs tested first as puppies and later again as puppies. Finally, there were no differences in consistency between working versus non-working dogs, between behavioral codings versus behavioral ratings, and between aggregate versus single measures. Implications for theory, practice, and future research are discussed.

Introduction

Personality in humans can be defined as the characteristics of individuals that describe and account for consistent patterns of feeling, thinking, and behaving [1]. Personality in nonhuman animals has often been defined operationally in terms of behaviors that are counted, timed, or rated in standardized tests [2]. In such cases, personality is defined in terms of 'correlated suites of behavior', where correlations in behavior can occur across different functional contexts, over time, or some combination of the two [3]–[7]. Personality assessments in species ranging from fish [8], arthropods [9], and cephalopods [10] to birds [11], hyenas [12], and nonhuman primates [13] meet a range of psychometric criteria held as standards for personality assessments in humans [14].

One of the first studies on dog personality was Nobel Laureate Ivan Pavlov's classic work on learning in which he classified dogs into four basic personality types based on their responses to conditioned reflex training [15]. The next significant contribution to personality research in dogs came 50 years later with Scott and Fuller's [16] work on genetic influences on dog behavior; as part of this work these authors attempted to predict later behavior in five different breeds of dogs from behaviors observed earlier in life. Later, in another study of behavioral prediction, Pfaffenberger [17] assessed dog personality using behavioral assessments (i.e., "puppy tests") to determine which puppies would be best suited for guide-dog work. Since these seminal studies, personality assessment has continued to flourish in the applied working-dog domain [18]–[19] and in studies of the nature and structure of dog personality itself [20]–[22], [23].

Dog personality research has been motivated, in part, by a number of practical concerns. First, there is widespread interest from potential companion dog owners in selecting dogs with personality characteristics that suit their lifestyle. Second, animal shelters and other agencies have an interest in using personality traits to improve the success of the adoption process and to direct care to the animals most in need of attention. Third, agencies focused on reducing the incidence of injuries caused by dog bites have an interest in identifying the individual animals most disposed to aggression. Fourth, working dog practitioners are interested in identifying the individuals with personalities most suited to successful job performance.

All of these cases rest on the assumption that behavior is at least somewhat consistent across time and/or situations. Thus, for example, dog owners try to understand how an adopted dog may behave once it is introduced to their home, or perhaps even months later, based on their observations and interactions with a dog in a shelter or a play area. Similarly, working-dog organizations are interested in identifying puppies as early in life as possible that will grow into adults suitable for working roles. Despite the long history of personality research in dogs and the considerable applied interests, to our knowledge, there are no published quantitative syntheses estimating the strength of consistency of dog personality. Instead, the dog-personality literature is populated with isolated studies in which personality is reported to be sometimes consistent and sometimes not (see Table 1). For example, one study [24] found a strong correlation (r=0.65) between biting/attack scores on two aggression tests held six months apart in adult dogs, and another study [25] found that dog personality assessments conducted as early as 8 weeks of age predicted suitability for police work at 18–24 months of age. Other studies have reported results that were dependent on the particular personality trait that was assessed. For example, one study found that shelter dogs recently adopted by new owners had strongly consistent separation anxiety across two weeks (r=0.71) but did not have strongly consistent fear of unfamiliar dogs (r =0.10) [26]. Still other studies have reported a lack of consistency for particular traits (e.g., fearfulness), or a decrease in consistency with increasing time intervals between tests [19], [27]. For example, one study reported differences in the consistency of dogs' confidence in semi-novel environments over shorter time intervals (e.g., 2–3 days, r=0.52) versus longer time periods (e.g., 1–3 months, r=0.16) [28]. In short, and as shown in Table 1, there is a lack of consensus about the extent to which personality is temporally consistent in dogs.

Table 1. A descriptive summary of the studies used in the meta-analysis.
https://doi.org/10.1371/journal.pone.0054907.t001

These inconsistencies in the literature have led several authors to conclude in qualitative reviews that 'puppy tests' are probably not worthwhile [23], [29]–[30]. In addition, there is a lack of consensus regarding the biological and measurement factors that influence personality consistency both in puppies and adult dogs. Here we attempt to quantify temporal consistency of personality in dogs and identify factors that may influence the strength of consistency. We do so by performing a meta-analysis on all known studies of dogs that provide relevant data.

The Nature of Dog Personality

There is some debate over the number and types of dimensions needed to characterize personality variation in most species of animals [4]. In humans, although not universally accepted, there is now considerable consensus that five broad personality dimensions can capture most of the behavioral variation in people [31]. Some attempts have been made to classify and conceptualize personality in nonhuman animals in terms of a five-dimension model [32]–[33], but there is little consensus about the number or type of personality dimensions that capture most of the observed behavioral variation in dogs.

Several models of dog personality derived from factor-analytic approaches have been proposed. For example, three- [22], [34] and five-dimension models [21] have been used to describe personality variation in companion dogs. One of the most widely used dog personality measures, the Canine Behavioral Assessment and Research Questionnaire (C-BARQ), proposes eleven different personality dimensions in companion dogs [20] and eight different personality dimensions in guide dogs [35]. These C-BARQ dimensions have demonstrated high levels of scale reliability and validity, and the instrument has been used in numerous studies across several populations of dogs [36]–[40].

Taking a different approach, Jones and Gosling [23] used an expert-sorting procedure to classify traits identified in all previous studies of dog personality. The procedure yielded seven broad dimensions (reactivity, fearfulness, activity, sociability, responsiveness to training, submission, aggression, as well as a classification for none/other, which included traits that were not considered to be personality).

These dog personality frameworks all share the goal of attempting to reduce the wide variety of behavior-descriptive terms into a more manageable set of broad dimensions, while also attempting to explain most of the personality variation in dogs. However, as even this brief synopsis illustrates, major differences still exist regarding the number and names of dimensions needed to characterize dog personality. Here we adopted the Jones and Gosling [23] framework, because it is applicable across all dog populations, and many studies examined in the current analyses were already classified in Jones and Gosling's expert-classification [23].

Potential Factors that Could Affect Personality Estimates

Numerous biological, environmental, and evolutionary influences could contribute to consistency estimates of personality in dogs. Here we focused on two broad areas relevant to temporal consistency in personality: Biological development and measurement methods.

A. Factors Related to the Biological Development of Dogs

Personality dimension.

Circulating hormone levels and hormone receptor density have been known to influence behavior [41]–[42]. Moreover, different endocrine systems are thought to have a larger influence on some behaviors than on others. For example, androgens are primary correlates of aggression [43]–[44], and corticosteroids often mediate stress, fear, and risk-taking behaviors [45]–[46].

If different personality dimensions are under different morphological or physiological constraints or if different dimensions undergo different rates of relative fixation throughout ontogeny, then we may expect there to be differences in the consistency estimates across personality dimensions. Personality dimensions that are correlated with developmental systems that are more sensitive to the environment should be less consistent through time than personality dimensions related to more stable developmental systems [47]. Past research does not afford clear predictions regarding which specific dimensions may be more or less consistent, but given the differences in constraints noted above, differences in consistency across different personality dimensions are possible, if not likely.

Age at first measurement.

In humans, personality tends to be more consistent in adulthood than in adolescence [48]–[50]. Studies of lifetime development of personality are rare in nonhuman animals, but this same pattern of increasing consistency with increasing age has also been observed in a lifetime study of squid [51]. There is some evidence that dogs exhibit more consistency of personality as they grow older [27], and significant personality consistency in adult dogs has been observed over an interval of one to two years [52]. Age-related patterns of personality consistency may result from the energetic or structural costs of changing one's personality, such as changes in neuroendocrine networks [53]. Similarly, increased age may result in increased personality consistency, if individuals choose ecological and social niches appropriate for their personality type, and these environments facilitate and encourage the expression of particular personalities [54]–[55].

Working versus non-working dog.

Many working-dog programs purpose-breed their own dogs, and dogs designated as breeders in these programs are often chosen based on their physical and behavioral characteristics, the latter of which often includes an individuals' propensity to consistently exhibit appropriate working behaviors. Also, large-scale working-dog programs tend to have standardized processes that could promote similarity that does not occur with non-working dogs [28], [56]. In addition, many of the dogs categorized in the non-working dog category here were sourced from a shelter environment, which is probably unstable and stressful [57] relative to rearing environments experienced by working dogs. Based on these observations, we expected working dogs to have higher estimates of consistent personality relative to non-working dogs.

B. Factors Related to Personality Measurements

Some factors that influence estimates of personality consistency may be related to the testing instrument; these factors could affect consistency estimates even if the actual underlying personality was stable. We focus on factors known to affect measurement in previous studies of human and nonhuman personality.

Interval between measurements.

Environmental influences such as changes in the social and developmental environment are more likely to occur during longer (versus shorter) intervals between tests. Some dog studies support the idea that shorter between-test intervals yield higher estimates of consistency than longer intervals [28].

Behavioral codings versus behavioral ratings.

At least two different types of methods can be used when measuring personality: behavioral codings and behavioral ratings [3], [58]. Behavioral codings typically attempt to measure observed, discrete classifications of behavior, often generated from an ethogram, such as the frequency and duration of a particular posture. For example, one study used the number of lines (marked on the floor of a test arena) that a dog crossed as an indicator of locomotor activity [59]. Ratings typically consist of broader judgments regarding a dog's standing on a behavioral trait made by people familiar with the dog. For example, in one study dogs were rated on a 1–5 Likert scale according to their playfulness with a rag in a standardized test [34].

In some instances, the reliability of behavioral ratings may be lower than that of behavioral codings because a rater has a restricted relationship to the target subject, such as in cases where raters only perform veterinary or feeding duties [60]. In other instances, empirical comparisons of the reliability of behavioral ratings versus behavioral codings of the same animals have shown behavioral codings to be less reliable than behavioral ratings [56]. This advantage of behavioral ratings can be attributed to the fact that behavioral ratings tend to reduce error variance by accounting for situational effects and incorporating longer behavioral trends. These factors would suggest that dog personality consistency estimates will be higher when behavioral ratings rather than behavioral codings are used.

Single versus aggregate measures.

Psychometric principles suggest that aggregate measures (i.e., sum or average of multiple observed behaviors) will tend to be more reliable than single measures because the random, nonsystematic error in the different multiple measures will tend to cancel each other out. In the human domain, aggregated measures yield greater consistency than do single measures [61]–[63]. In dogs, aggregate measures have been shown to be more powerful predictors than single measures with regards to military working-dog certification outcomes [28].

Similarity of tests.

In many studies, two different test methods are used to measure the same personality trait. For example, at one time point, behavior may be measured using a standardized test situation and behavioral coding (e.g., shelter tests for behavior) but at a second time, a rating form might be used [64]. Studies that used the same test to measure a certain personality dimension are likely to yield higher consistency estimates than studies that used different tests across time, due to reduced method variance in the former [65].

Summary.

A great deal of research has accumulated on the temporal consistency of dog personality. Here we use meta-analytic methods to quantitatively summarize the overall consistency of dog personality and the factors that influence it. In line with our synopsis above, we expected our meta-analysis to reveal that dog personality was moderately consistent over time, but the absolute level of consistency may vary depending on the personality dimension being assessed. Further, we predicted that personality will likely be more consistent when dogs are tested as adults, when dogs are sourced from working dog programs, when behavioral ratings are used, when aggregate measures are used, when shorter test intervals are used, and when the same test is administered across test occasions.

Methods

Literature Search Procedures

To identify as many relevant studies as possible, we first searched *PsychInfo, Biosis, Web of Science*, and *ProQuest Dissertation and Theses* electronic databases for documents catalogued before August 2011. In each database, 12 keyword searches were performed, derived from all combinations of a 6 (temperament test, personality test, behavior evaluation, prediction, temperament, personality) x 2 (dog, puppy) matrix. Next, we examined the reference sections of all of articles obtained through the database searches to determine if any cited works had titles that also might be relevant to the topic. Also, we conducted a Social Science Citation Index search on two previous studies that were heavily cited in the topic of dog personality [18], [23]. Finally, to reduce bias potentially introduced by our limited perspective we also asked eight dog-personality experts if they knew of any studies we had missed.

Inclusion Criteria

For a study to be included in the meta-analysis, several criteria had to be met. First, studies had to have tested the consistency of dog behavior across two or more time points. Second, studies had to provide a bivariate correlation (*r*) of this relationship or provide sufficient information to compute or convert the estimates to an *r*-value. Third, to ensure that all estimates met a threshold of peer-reviewed or committee-reviewed scholarship standards, studies had to be either a published report or a completed dissertation or thesis. Finally, to ensure that the estimates were assessing repeated measurements of the same construct, the second test assessment method had to be at least conceptually similar (i.e., measure the same dimension) as the first.

Our search yielded 107 studies. Each was examined by the first or second author and 79 studies were excluded based on our selection criteria. The most common causes of exclusion were because the studies presented data in a way that was impossible to convert to a bivariate correlation [66] or because the test two measurements did not clearly measure the same dimension at two different time points [67]. In 19 cases where the article did not report sufficient information to compute *r*, but data from the study was relevant and recent (i.e., published in 2000 or later), we emailed the study's authors to request the required information. Five author groups responded, and three were able to provide the additional relevant data. In total, 31 studies (28 published studies and 3 unpublished dissertations) with a total of 822 estimates of consistency were included in the meta-analysis.

Information Coded from Studies

For each study, we recorded information regarding authors, year, title, journal, source we found each study from, the number of subjects, the age of the subjects at the first test in weeks (if more than one age was given for a single estimate, we used the average age), the type of dogs (working or non-working dog), the name of the trait(s) given from the study, a description of the test domains (e.g., response to novelty), the average interval between tests (in weeks), the test methods (codings or ratings), if the measures of personality reflected a single measure or an aggregate set of measures, and if test one was exactly the same or only conceptually the same as test two.

The studies amassed here reported 213 unique trait names, which precluded any comparison of consistency estimates at the trait level due to a lack of sufficient statistical degrees of freedom. Therefore, to compare traits assessed in the studies, we classified all trait names given in each study in terms of the personality dimensions described in the Jones and Gosling [23] seven-dimension framework (see Table 2). For studies that were not included in the Jones and Gosling [23] review, we used descriptions of trait names, test domains, and test procedures given by authors to match them to traits already classified by Jones and Gosling [23] (see Table 3). Jones and Gosling [23] questioned the wisdom of separating the fearfulness and reactivity dimensions and only one study in our review provided estimates that could potentially be classified exclusively as 'reactivity' [68] so we combined these two dimensions, using the label 'fearfulness.' We also excluded descriptions of traits considered as 'other', because Jones and Gosling [23] considered these traits as not being related to dog personality. In cases where traits from studies were identified as straddling more than one of the seven dimensions, we assigned the trait to both categories (see Table 3).

Table 2. Description of the Jones & Gosling 2005 personality framework used for meta-analysis.
https://doi.org/10.1371/journal.pone.0054907.t002

Table 3. Classification of study traits into the Jones and Gosling (2005) 7-dimension dog personality framework.
https://doi.org/10.1371/journal.pone.0054907.t003

We recorded several additional characteristics, including the analysis strategy that authors used to determine the relationship between tests (e.g., correlation), the result type (e.g., odds ratio or Pearson's r, etc.), the direction of the effect, and the value of r. In studies that reported results from logistic regression, we converted log odds ratios to Cohen's d, and then converted d to r [69]. For studies that reported only a p-value, we converted p to its associated one-tailed standard z and then converted z to r [70]. If a study reported a test-retest reliability coefficient based on opposite response scales across tests (e.g., higher numbers indicated greater presence of a trait on the response scale for the first test while higher numbers indicated lower presence of the trait on the response scale for the second test), we re-keyed the coefficient.

Data Reliability

The second author recorded information for all studies, and the first author independently recorded data from 10 of the studies to assess reliability. There was 96% agreement between the two authors for those 10 studies across 22 variables. The few disagreements were about how to classify the original personality trait name from the study into the Jones and Gosling [23] framework; so for personality dimension classifications, the first author also categorized trait names from all studies according to the Jones and Gosling [23] framework. Discrepancies were noted and discussed, and agreement was reached in all cases.

Methods of Data Integration

Before conducting any analyses, we examined the distribution of effect sizes to determine if our dataset contained statistical outliers. Grubbs' test was applied [71]–[72] and no outliers were identified.

We employed Duval and Tweedie's [73]–[74] trim-and-fill procedure to test whether the distribution of effect sizes used in the analyses was consistent with that expected if the estimates were normally distributed. If the distribution of observed effect sizes was skewed, indicating a possible bias created either by the study retrieval procedures or by data censoring on the part of authors, the trim-and-fill method provides a way to estimate the values from missing studies that need to be present to approximate a normal distribution. The procedure then imputes these missing values, permitting an examination of an estimate of the impact of data censoring on the observed distribution of effect sizes.

Fixed and random error.

There are two common ways to conceptualize meta-analysis: fixed effects and random effects models. These models differ in their theoretical assumptions and also how mean effect sizes and significance are calculated [75]. Fixed error models assume that studies in a meta-analysis are sampled from a single population with a fixed` average effect size. Random effect models assume that the average effect size varies randomly from study to study: studies in a meta-analysis come from multiple populations that have different average effect sizes, so study/population effect sizes can be thought of as being sampled from a 'superpopulation' [76]. One consequence of these assumptions lies in the statistical calculation of error. Fixed error models assume error is introduced because of sampling studies from a population of studies. This error is also assumed in random effects models, but in addition there is error created by sampling the populations from a superpopulation [77]. Fixed effect models are common but there is considerable theoretical and empirical evidence that real-world data likely fit random effects models more closely [77]–[78]. For this reason, we reported average effect sizes computed under the random effects model, but conducted our analyses using the fixed effect model too. In fact, random effects are computed in an iterative fashion based on fixed effects. We report within-class goodness-of-fit values (Q_w) using fixed-effect weights to assess the between-studies dispersion and tests of homogeneity for all studies within a group. Q_w values computed using random-effects weights would not be appropriate for this purpose. Average effects computed using fixed-effect weights and other fixed effect model results are presented in the supporting information.

Calculating average effect sizes.

A weighting procedure was used to calculate average effect sizes, both within and across studies. Each independent correlation was first multiplied by the inverse of its variance. Then, the sum of these products was divided by the sum of the inverses [79]. This weighting procedure is generally preferred because it gives greater weight to effect sizes based on larger samples since larger samples give more precise population estimates. Also, 95% confidence intervals were calculated for the overall weighted average effect. If the confidence interval did not contain zero, then the null hypothesis that there is no consistency in dog personality across time was rejected.

Identifying independent hypothesis tests.

One problem that arises in calculating effect sizes involves deciding what constitutes an independent estimate of effect. Here, we used a shifting unit of analysis approach [80]. In this procedure, each effect size associated with a study is first used as if it were an independent estimate of the relationship. For example, if a single study provided correlations across two time points for both aggression and fearfulness, two separate correlations were recorded. However, for estimating overall consistency in dog personality, these two correlations were averaged prior to analysis, so that the study only contributed one effect size. To calculate the overall weighted correlation and confidence interval, this one effect size would be weighted by the inverse of its variance and sample size. However, in an analysis that examined consistency in aggression and fearfulness separately, the study was permitted to contribute one effect size to each mean effect size. The shifting unit of analysis approach retains as much data as possible from each study while holding to a minimum any violations of the assumption that data points are independent.

Tests for moderators of effects.

Possible moderators of the consistency of dog personality between two time points were tested using homogeneity analyses and meta-regression techniques [79]–[81]. Homogeneity analyses compare the amount of variance in an observed set of effect sizes with the amount of variance that would be expected by sampling error alone. The analyses can be carried out to determine whether the variance in a group of individual effect sizes varies more than predicted by sampling error. Within a fixed effects model, the

homogeneity of the set of effect sizes is tested using a within-class goodness-of-fit statistic (Q_w), which has an approximate chi-square distribution with $k - 1$ degrees of freedom, where k equals the number of effect sizes. Thus, a significant Q_w statistic would indicate systematic variation among effect sizes and suggest that moderator variables be examined or that a random effects model may be most appropriate for the data [80]. Homogeneity analyses can also be used to determine whether multiple groups of average effect sizes vary more than predicted by sampling error alone. In this case, statistical differences among groups of estimates are tested by computing the between-class goodness-of-fit statistic (Q_b), which has a chi-square distribution with $p - 1$ degrees of freedom, where p equals the number of groups. A significant Q_b statistic indicates average effect sizes vary between categories of the moderator variable more than predicted by sampling error alone. This strategy is analogous to testing for group mean differences in an analysis of variance.

Meta-regression techniques were used when moderators were continuous (i.e., interval between tests measured in weeks). Two values assess the fit of a weighted regression model. First, the Q regression (Q_r) examines the total variability associated with the predictors in the regression model. Q_r has p degrees of freedom, where p equals the number of predictors. A significant Q_r indicates the regression model explains significant variability in effect sizes and that at least one regression coefficient is significantly different from zero. The weighted sum-of-squares residual (Q_e) examines the variability unaccounted for by the model. Q_e has $k - p - 1$ degrees of freedom, where k represents the number of effect sizes and p equals the number of predictors. A significant Q_e indicates that after removing variability based on the predictor values, the effect sizes remain heterogeneous [79]. While in a fixed effect model, a significant Q_e may suggest that a random effects model may be more appropriate for data, in a random effects model, this residual heterogeneity is assumed to be composed entirely of sampling error and will generally be small [79]. Meta-regression is analogous to testing for effects of a set of predictors on an outcome variable in a multiple regression model. For this study, we performed an unrestricted maximum likelihood (ML) random effects regression, using the consistency correlation as the dependent variable.

All analyses were conducted using Comprehensive Meta-Analysis statistical software package Version 2 [82].

Results

Table 1 summarizes descriptive information from the 31 studies included in our analyses. 96.8% of the studies reported significant consistency in at least one trait, 67.7% of studies reported a null effect in at least one trait, and 64.5% of studies reported both a significant and non-significant effect in at least one trait. The sample size of each study varied considerably, ranging from 7 dogs [59] to 938 dogs [35], with an average of 84 dogs per study. The time interval between measures across studies varied as well, ranging from 3 days [83] to 224 weeks [37], with an average interval of 21 weeks. The age at which a dog was first tested also varied considerably across studies, ranging from 3 weeks of age [84] to 296 weeks of age [85], with a mean of 49 weeks of age. 65% of dogs tested were non-working dogs. For working dog studies, 6 studies surveyed guide dogs, 1 study surveyed military working dogs, 1 study surveyed police dogs, and 1 study surveyed detector dogs. Four studies did not specify what type of dog (working or non-working) was being tested. A majority (72%) of studies used behavioral ratings at both test periods, but of these, a majority (76%) of the raters did not own or care for the dog. More studies (59%) used aggregate rather than single measures at both test periods. More studies (57%) used the exact same measure at both time points than used conceptually similar, but different measures.

Overall Consistency in Dog Personality

Of the 822 effect sizes, 708 were in a positive direction, and 114 were in a negative direction. The effect sizes for estimates of consistency for single traits ranged from $r=-0.73$ [40] to $r=1.00$ [86]. The overall effect size was moderate and significantly greater than zero ($k=31$, $r=0.43$, 95% CI $=0.35$, 0.50). Additionally, the tests of the distribution of effect sizes revealed that we could reject the hypothesis that the effects were estimating the same underlying population value (fixed effects: $Q_{30} = 3393.30$, $p<0.001$), suggesting that we could proceed with moderator analyses and that a random effects model is likely most appropriate. Trim-and-fill analyses using random error models, explicitly searching for possible missing effects on the left side of the distribution (those that would reduce the size of the positive average r) revealed no evidence for possible data censoring.

Moderators of Consistency Estimates

Personality dimension.

Personality traits that could not be clearly classified into a single Jones and Gosling [23] dimension were not used in the moderator test (i.e., all estimates used fell into only one dimension). Using the six dimensions from Jones and Gosling's [23] framework and including dogs across all ages, the average weighted consistency correlation across time points did not vary significantly by dimension ($Q_5 = 6.60$, $p=0.25$). Consistency estimates for all personality dimensions were significantly different from zero, ranging from 0.28 for responsiveness to training to 0.50 for aggression (Table 4).

Table 4. Results of moderator analysis.

Age at first measurement.

The association between the age of the dog at first test and the personality consistency estimate was assessed by categorizing estimates according to whether dogs were puppies (<12 months old) or adults (>12 months old) at the first test. For both age categories, consistency estimates were significantly different from zero and significantly different from each other ($Q_1=6.58$, $p=$

0.01; Table 4). Of note, the average weighted adult personality consistency estimate ($r=0.51$) was 1.7 times as large as the puppy personality consistency estimate ($r=0.30$).

To test the effectiveness of 'puppy tests' more explicitly we also examined whether consistency estimates were different between puppies first tested as puppies and then tested again as puppies (average interval between tests =7.84 weeks) versus puppies first tested as puppies and then tested as adults (i.e., average interval =47.52 weeks). For both categories, consistency estimates were significantly different from zero ($r=0.38$ and 0.40, respectively), but were not different from one another ($Q_1=0.02$, $p=0.90$).

Next, we examined whether consistency varied by personality dimension separately for puppies and adult dogs (Table 5). Among puppies, the estimates for all dimensions except for sociability were significantly different from zero, ranging from 0.16 for responsiveness to training to 0.51 for aggression. Further, this variability in consistency by personality dimensions was significantly greater than would be expected by sampling error alone ($Q_5=22.12$, $p<0.001$). A series of pair-wise comparisons suggested the largest effect sizes for consistency in puppies were for aggression ($r=0.51$) and submissiveness ($r=0.43$), which were not significantly different from each other. Fearfulness ($r=0.24$) and responsiveness to training ($r=0.16$) were the least consistent dimensions and were not significantly different from each other. Responsiveness to training and fearfulness were significantly less consistent than aggression and submissiveness but not activity. Activity ($r=0.26$) was significantly less consistent than submissiveness and marginally less significant than aggression (Table 6).

Table 5. Interaction of personality dimensions and age of dog.
https://doi.org/10.1371/journal.pone.0054907.t005

Table 6. Contrasts (Q-values) for personality dimensions in puppies.
https://doi.org/10.1371/journal.pone.0054907.t006

In contrast, there was no significant variation in consistency by personality dimension among adult dogs, $Q_5 =2.70$, $p=0.75$ (Table 4). Rather, with the exception of a non-significant consistency estimate for submissiveness ($r=-0.13$), consistency estimates among dogs for all other personality dimensions were significantly different from zero and were fairly similar, ranging between 0.47 for sociability to 0.51 for fearfulness. It should also be noted that the estimate for the submissiveness dimension among adult dogs was based on only a single study and should be interpreted with caution.

Working versus non-working dogs.

Consistency estimates for both working and non-working dogs were significantly different from zero ($r=0.36$ and 0.41, respectively), but there was no difference in consistency of dog personality between the two groups ($Q_1=0.29$, $p=0.59$; Table 4).

Interval between measurements.

To evaluate the association between test interval and consistency, we used the test interval (in weeks) as a continuous variable in an unrestricted ML meta-regression. The model was significant ($Q_1 =13.57$, $p<0.001$), with ~2% of the variability in consistency correlations accounted for by test interval. As the test interval increased, the magnitude of consistency decreased (Table 7).

Table 7. Unrestricted ML meta-regression for 'time interval between tests' moderator.
https://doi.org/10.1371/journal.pone.0054907.t007

To further investigate the effectiveness of 'puppy tests', we examined the association between test interval and age of dog at first test on the overall effect size of consistency. We separated test interval into short (<10 weeks), medium (10–24 weeks), and long (>24 weeks) categories and used the previous designations for puppies (<12 months old) and adults (>12 months). For puppies, estimates for all interval categories were greater than zero and there was no difference between length-of-interval category ($Q_2 = 1.586$, $p=0.45$). For adults, all estimates for interval categories were greater than zero and there was a marginal statistical difference between interval categories ($Q_2=4.46$, $p=0.096$). Short intervals ($r=0.60$) tended to result in higher consistency estimates than long intervals ($r=0.32$; $Q_1=4.69$, $p=0.03$). There was no difference between short and medium intervals ($Q_1=0.25$, $p=0.62$), or between medium and long intervals ($Q_1=1.38$, $p=0.24$; Table 8).

Table 8. Interaction of test interval and age of dog.
https://doi.org/10.1371/journal.pone.0054907.t008

The consistency estimates for both behavioral codings ($r=0.42$) and ratings ($r=0.42$) were significantly different from zero, but the two estimates were not significantly different from one another ($Q_1=0.00$, $p=0.99$; Table 4). Note that for this moderator test we used only studies that used the same method of measurement on both test occasions (e.g., behavioral codings or ratings at both test 1 and test 2, but not a combination of the two methods across time).

Consistency estimates for aggregate trait measures ($r=0.45$) and single trait measures ($r=0.40$) were significantly different from zero, but not different from one another ($Q_1=0.55$, $p=0.46$; Table 4). Note that for this moderator test we used only studies that used the same measure for both tests (e.g., single measures or aggregate measures at both test 1 and test 2).

There was a significant difference in consistency of dog personality when the exact same test was given across time points compared to when the test was methodologically different ($Q_1=11.72$, $p=0.001$; Table 4). Consistency was 1.8 times greater when the tests across the two measurement time points were identical ($r=0.49$) compared to when different tests were given ($r=0.27$).

Discussion

The concept of personality implies that behavior shows some level of temporal consistency. In dogs, personality consistency is especially relevant because success in most tasks depends on a dog's ability to express predictable and appropriate behaviors. Canine researchers [22]–[23] widely recognize the existence of personality in dogs, but there has been little clarity regarding the nature and strength of personality consistency and the usefulness of 'puppy tests' in predicting adult behavior [29]–[30]. This lack of clarity is evidenced by a vote-count of prior results of personality consistency in dogs (Table 1), where 64.5% of studies report both positive and negative findings for personality consistency. Using meta-analysis, we quantitatively synthesized previous results to determine when personality may be consistent, the factors that influence personality consistency, and to provide recommendations regarding studies that will likely further the field.

Our results provide evidence for the broad proposition that dog personality is moderately consistent (overall average weighted effect size $r=0.43$). This finding fits well with findings on the consistency of behavior in non-domesticated animals, where a recent meta-analyses reported an overall average weighted effect size of $r=0.37$ [47] using fixed effects models (our fixed effect estimate was $r=0.29$; see supporting information) for correlations of the same behavior through time. Another meta-analysis reported an overall average weighted effect size of 0.20 [87] using random effects models for correlations between different behaviors at the same time (i.e., a behavioral syndrome).

One factor that influenced personality consistency in dogs was age. Average weighted personality consistency estimates were different from zero in both puppies ($r=0.31$) and adults ($r=0.51$) and the adult dog personality consistency estimate was significantly greater than that observed in puppies. These results are in line with previous findings in dogs [27] and humans [49]. From a developmental perspective, increasing consistency or predictability may be observed with age if there are energetic or structural costs to changing one's personality [53]. Similarly, some theory suggests that strong personality consistency is to be expected when positive feedback loops exist between the individual and its environment [55], such as when individuals prefer to live in environments in which they perform at high levels [55], [88]. Most dogs do not select the environment in which they live in but dynamic social interactions with humans occur throughout ontogeny and could also in theory reinforce behaviors deemed appropriate by owners and other handlers. In theory, aging would allow for positive feedback loops to have a greater effect in adults than in puppies.

Another factor that influenced personality consistency in dogs was the interaction between personality dimension and age. In puppies, aggression and submissiveness were the most consistent dimensions, and neared estimates of personality consistency found in adults, while responsiveness to training and fearfulness were the least consistent dimensions (sociability consistency estimates in puppies was not different from zero). In adult dogs this was not the case; instead, all dimensions were equally consistent with the exception of submissiveness, where our analysis was restricted to a single study. Unfortunately, little is known concerning relative rates of development of different personality dimensions in puppies. However, one putative explanation for the observed differences in consistency among different personality dimensions in puppies may be proximate hormonal mechanisms. For example, androgens are known to influence both aggression [44] and submission [89] in other taxa. Corticosteroids are known to influence fearfulness [45]–[46]. Structurally, if different personality dimensions tend to be influenced by different proximate mechanisms, and these different underlying mechanisms have different rates of physical ontogeny (e.g., organ development, receptor density development, etc.), then different personality dimensions could end up having different rates of 'fixation', or observed consistency through time. From a practical perspective, it is worth noting that responsiveness to training was one of the least consistent personality dimensions observed in puppies, despite its importance to the general public [90]. These results imply that snapshots of responsiveness to training in puppies may not be an accurate assessment of a puppy's ability to respond or learn later training or obedience exercises [19]. Further studies are needed on the factors that influence consistency in all personality dimensions, but perhaps in particular responsiveness to training in puppies, because this dimension is especially relevant to human-dog relationships.

Time interval between tests was also found to play a small, but significant role on personality consistency estimates, with an overall negative relationship being observed. However, this effect of decreasing consistency with increasing interval was found mainly in adult dogs, where shorter (<10 weeks) time intervals tended to result in larger consistency estimates ($r=0.60$) than longer (>24 weeks) time intervals ($r=0.32$). All categorical time interval estimates were different from zero in adults but estimates of consistency

fell from $r=0.60$ over short intervals to $r=0.32$ for long intervals. These smaller effect sizes for adult personality consistency over longer periods of adult life suggest that even as adults, personality dimensions are not fixed properties of individual dogs, but may instead be conducive to environmental and social manipulation and change.

Puppy consistency estimates for short, medium, and long test intervals were not different from one another, but were different from zero ($r=0.25 - 0.39$). In addition, we found no difference in personality consistency estimates between dogs first tested as puppies and then either as puppies or adults during the second test period ($r=0.38$ versus $r=0.40$, respectively). One of the core questions facing many working and companion dog organizations is whether 'puppy tests' are predictive of later adult behavior [30]. Our results suggest that puppy personality is moderately consistent, and remains so, throughout the juvenile and into the adult period. This may especially true for particular personality dimensions, such as aggression or submissiveness, which appear to be as consistent as dimensions measured in adult dogs. Our results suggest that the blanket idea that 'puppy tests do not work' needs to be reconsidered – it may depend both on the personality dimension being considered as well as factors individual dogs experience throughout their life such as litter size [91]–[92], body mass, and early growth [93]. As in adults, puppy personality can be characterized as being both moderately consistent as well as sometimes highly plastic, depending on the personality dimension of interest.

We did not find a difference in personality consistency in working versus non-working dogs. We examined the possibility that (in)stability of the rearing environment might alter the consistency of personality, but we did not find evidence to support this idea one way or the other. Other studies have reported that some dog breeds have more consistent personalities than others [94], and it may be that consistency itself is also a 'dimension' that could be selected for, as opposed to selecting for absolute levels of behavior that are observed at any one time point. To our knowledge, no professional working dog programs or companion dog breeders have yet to investigate the possibility of selection for consistency per se, but our results suggest that personality consistency has not yet been altered by any artificial selection imposed by working dog programs.

Somewhat surprisingly, we did not find any differences in consistency estimates between behavioral codings versus behavioral ratings and between single versus aggregate measures. In principle, behavioral ratings and aggregate measures should yield more consistency because in both cases, error variance is reduced [61]. One explanation for this result may lie in the close social relationships dogs have shared with humans over the past 14,000 years [95]. In dogs, as opposed to other nonhuman animals, this relationship may have resulted in accurate perception of dog behavior by humans, regardless of the measurement method (i.e., error variance in coding methods and single measures are not different from ratings methods and aggregate measures). An alternative explanation somewhat supported by our results is that behavioral ratings yielded lower estimates of consistency than expected because observers using ratings methods had restricted relationships to the target subjects [60]. For 38% of the studies we were unable to determine the exact nature of the relationship between the rater and the subject but only 24% of studies that used ratings were from studies where longer-term knowledge of the dog could be implied (i.e., owners gave dogs' ratings). This issue reflects the current state of the literature (dog personality research is dominated by studies using working dogs [23], which use program staff with unknown personal relationships to the dogs to provide ratings).

It is encouraging that personality consistency estimates were moderate both when using single trait measures and aggregate trait measures. Again, this pattern was somewhat unintuitive, based on measurement theory. Our results could be explained if the close social ties between humans and dogs resulted in the researcher's ability to define single behavioral indicators (such as particular postures) that were precisely recognized by observers and that have strong ties to broader personality patterns. Alternatively, our results could have also occurred if aggregate measures consisted of behaviors that did not cohere and so should not have been combined together, thus decreasing aggregate measures' predictive validity [96]. Some analyses suggest that behaviors that are theoretically part of the same behavioral category may not be empirically related [56]. Unfortunately, dog personality studies often do not report measures of internal coherence of aggregate scales [96], so we were unable to test the idea that reliable aggregate measures had different consistency estimates than aggregate measures with unknown or low internal coherence reliability estimates. Nonetheless, the current results indicate that the choice of measurement method (i.e., behavioral codings versus behavioral ratings and single versus aggregate measures) may not be critical when determining the best way to measure consistent behavioral properties in dogs.

In principle, the greater the similarity between tests administered at different time points, the more consistent dog personality should appear because method variance is reduced. Our meta-analysis provided strong support for this pattern, showing that consistency was greater when testing instruments were identical across time points compared to when the two tests differed. Practical concerns drive many researchers to use different tests. For example, it is convenient for shelter staff to give behavioral assessments to dogs while in the shelter but exact follow up tests are not possible and instead are often conducted using a questionnaire that are given to the adoptive owner at a later date [e.g., 26, 59]. When these methods are used and the same behavior is measured in two different ways, personality may appear to be less consistent but the consistency estimate will be confounded with method variance. Of course, there are times when the same test cannot be given but our analyses suggest that efforts to create tests that are as conceptually similar as possible would be worthwhile. One potential issue with this moderator is that when the type of test differs between first and second assessments, so does the testing context. Thus, test type (e.g., exact same versus conceptually same behavioral assessments) is partially confounded in our analyses with test context (e.g., at a shelter vs. in the home). Future research is needed to separate the effects of these different factors.

Limitations and Recommendations.

Despite the obvious importance of understanding the consistency of personality in dogs, there is currently a paucity of studies examining the factors that influence personality consistency. Indeed, many pertinent research questions could not be addressed due to the small number of samples available for moderator analyses. For example, many dog personality studies focus on how well an earlier behavioral test can predict later 'success' or certification in a training program but 'success' is usually not well-defined (i.e., usually reported as a yes/no outcome) [97]. Attempts at defining explicit domains of success/failure (e.g., quantitative descriptions of behavior) could enable potential inclusion of these studies in future meta-analyses. Also, there could be differences

based on training methods, but there were not enough studies to examine training differences based on different types of programs, shelters, or even by country. Finally, it would have been interesting to explore the role of moderators in a hierarchical fashion or simultaneously, but the lack of degrees of freedom limited our ability to do so.

Surprisingly, many studies did not even report the breeds used or individual breed results. There are differences in absolute levels of personality expression between breeds and breed clusters [98], so it is possible there are breed differences in personality consistency too. In addition, potentially important early environmental factors often go unreported. For example, across the 31 studies we examined, only 8 reported the weaning age of the dog and only 6 reported the age in which dogs were first housed singly. Early experiences, such as exposure to novel stimuli and socialization, are important factors in the biological development of dogs [99]–[100]. From a statistical standpoint, many studies did not report consistency in ways that could be translated to effect sizes (e.g., only mean level results were reported [101]) and many studies did not report confidence intervals. Effect size estimates and confidence intervals allow one to easily identify the relationship between tests and compare results across different samples, measures, and conditions [102]. We recommend effect sizes (r or d) and 95% confidence intervals are reported in all future dog personality studies. Changes in reporting practices could do much to alleviate this issue.

Currently, there are very few studies on personality development patterns. Just as some personality dimensions may be more or less consistent than others, some individuals' personalities may also be more or less consistent. Evidence from other taxa suggest that particular personality 'types' may be more or less capable of altering their behavior appropriately to environmental conditions [51], [103], and these personality by personality plasticity interactions may be widespread [104]. Indeed, the consistently moderate effect sizes for personality consistency witnessed here could be explained if some individuals were strongly consistent in their personality expression while others were not. Individual differences in consistency per se have important implications for a practical understanding of personality development and human-dog relationships; to our knowledge no dog studies have explicitly examined this aspect of personality development.

Little is known concerning the relationship between the development of proximate biological mechanisms (i.e., neuroendocrine systems) and the development of personality consistency in dogs. Some underlying biological mechanisms may have different rates of biological development [30]. There is a clear need to understand how differential rates of development of proximate mechanisms may impact the consistency of different personality dimensions. Likewise, studies on environmental effects that differentially affect personality dimension expression would be equally useful.

There are three other important limitations of any meta-analysis. First, moderator variables are likely confounded with one another in complex ways that make it difficult to tease apart the independent effects of each moderator. Second, it is possible that factors that are not examined also explain variability in effect size, perhaps better than those chosen in this meta-analysis. Third, synthesis-generated evidence should not be interpreted as supporting statements about causality [102]. When groups of effect sizes are compared within a research synthesis, the synthesis can only establish an association between a moderator variable and the outcomes of studies, not a causal connection. It is important that future research examine the factors that this meta-analysis has identified as influential in explaining differences in dog personality consistency using more controlled designs.

Summary.

Taken together, our results indicate that personality is generally consistent in dogs. The question of personality consistency in dogs has implications for many areas of human society, and our meta-analysis is a first step towards quantitatively synthesizing the existing information on personality consistency in dogs. Along with the general theme of consistency observed here, some of the important factors that tended to influence consistency estimated in dogs include age, personality dimension, test interval, and the conceptual similarity between test situations. In puppies, the predictive validity of 'puppy tests' is most likely to be detected when measuring aggression and submissiveness and less so in other personality dimensions. These latter personality dimensions (responsiveness to training, fearfulness, activity, and sociability) may be more amenable to analyses of how, why, and when personality changes. In adult dogs, personality consistency was stronger than in puppies and equally predictable across all dimensions examined. Adult personality consistency estimates improved with a decreasing time interval between tests (which was not the case in puppies), but over longer time periods, personality in adults could be described as being moderately plastic. Our results suggest that useful future studies could examine the developmental rates of different proximate mechanisms underlying different personality dimensions, to address whether consistency per se can be viewed as a dimension in and of itself, and to identify the specific periods of life during which different personality dimensions stabilize. In addition, improved reporting methods are urgently needed to furnish researchers, working-dog organizations, breeders, shelters, and pet owners with the tools required to identify the factors likely to be responsible for personality predictability and change.

Supporting Information

Figure S1.
PRISMA 2009 Flow Diagram.
https://doi.org/10.1371/journal.pone.0054907.s001
(DOC)

Supporting Information S1.
Results of Moderator Analysis using fixed effects models.
https://doi.org/10.1371/journal.pone.0054907.s002
(DOC)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JF DLS SDG. Performed the experiments: JF DLS. Analyzed the data: JF EP. Contributed reagents/materials/analysis tools: EP. Wrote the paper: JF DLS EP SDG.

References

1. Pervin LA, Cervone D, John OP (2005) Personality: Theory and Research (9th ed.). Hoboken: Wiley. 640 p.

2. Dingemanse NJ, Wolf M (2010) Recent models for adaptive personality differences: a review. Phil Trans R Soc B 365: 3947–3958.
   View Article • Google Scholar

3. Gosling SD (2001) From mice to men: What can we learn about personality from animal research? Psychol Bull 127: 45–86.
   View Article • Google Scholar

4. Gosling SD (2008) Personality in non-human animals. Soc Personal Psychol Compass 2: 985–1001.
   View Article • Google Scholar

5. Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ (2007) Integrating animal temperament within ecology and evolution. Biol Rev 82: 291–318.
   View Article • Google Scholar

6. Sih A, Bell A, Johnson JC (2004) Behavioral syndromes: an ecological and evolutionary overview. Trends Ecol Evol 19: 373–378.
   View Article • Google Scholar

7. Sih A, Bell AM, Johnson JC, Ziemba RE (2004) Behavioral syndromes: An integrative overview. Q Rev Biol 79: 242–277.
   View Article • Google Scholar

8. Budaev SV (1997) 'Personality' in the guppy (Poecilia reticulata): A correlational study of exploratory behavior and social tendency. J Comp Psychol 111: 399–411.
   View Article • Google Scholar

9. Wray MK, Matilla HR, Seeley TD (2011) Collective personalities in honeybee colonies are linked to colony fitness. Anim Behav 81: 559–568.
   View Article • Google Scholar

10. Sinn, D L., Perrin, N A., Mather, J A., & Anderson, R C. (2001). Early temperamental traits in an octopus (Octopus bimaculoides). Journal of Comparative Psychology, 115, 351–364.

11. Carere C, Drent PJ, Privitera L, Koolhaas JM, Groothuis TG (2005) Personalities in great tits, Parus major: stability and consistency. Anim Behav 70: 795–805.
    View Article • Google Scholar

12. Gosling SD (1998) Personality dimensions in spotted hyenas (Crocuta crocuta). J Comp Psychol 112: 107–118.
    View Article • Google Scholar

13. Pederson AK, King JE, Landau VI (2005) Chimpanzee (Pan troglodytes) personality predicts behavior. J Res Pers 39: 534–549.
    View Article • Google Scholar

14. Gosling SD, Vazire S (2002) Are we barking up the right tree? Evaluating a comparative approach to personality. J Res Pers 36: 607–614.
    View Article • Google Scholar

15. Pavlov IP (1941) Lectures on conditioned reflexes, vol. 2: Conditioned reflexes and psychiatry. W. Horsley Gantt, Translator. New York: International Publishers.

16. Scott JP, Fuller JL (1965) Genetics and the social behavior of the dog. Chicago: University of Chicago Press. 506 p.

17. Pfaffenberger CJ (1976) Guide dogs for the blind: Their selection, development and training. Amsterdam: Elsevier. 238 p.

18. Goddard ME, Beilharz RG (1984) A factor analysis of fearfulness in potential guide dogs. Appl Anim Behav Sci 12: 253–265.
    View Article • Google Scholar

19. 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.
    View Article • Google Scholar

20. Hsu Y, Serpell JA (2003) Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J Am Vet Med Assoc 223: 1293–1300.

View Article • Google Scholar

21. Ley J, Bennett P, Coleman G (2008) Personality dimensions that emerge in companion canines. Appl Anim Behav Sci 110: 305–317.
View Article • Google Scholar

22. Svartberg K, Tapper I, Temrin H, Radesäter T, Thorman S (2005) Consistency of personality traits in dogs. Anim Behav 69: 283–291.
View Article • Google Scholar

23. Jones AC, Gosling SD (2005) Temperament and personality in dogs (Canis familiaris): A review and evaluation of past research. Appl Anim Behav Sci 95: 1–53.
View Article • Google Scholar

24. Netto WJ, Planta DJU (1997) Behavioural testing for aggression in the domestic dog. Appl Anim Behav Sci 52: 243–263.
View Article • Google Scholar

25. Slabbert JM, Odendaal JSJ (1999) Early prediction of adult police dog efficiency–a longitudinal study. Appl Anim Behav Sci 64: 269–288.
View Article • Google Scholar

26. Stephen J, Ledger R (2007) Relinquishing dog owners' ability to predict behavioural problems in shelter dogs post adoption. Appl Anim Behav Sci 107: 88–99.
View Article • Google Scholar

27. Goddard ME, Beilharz RG (1986) Early prediction of adult behaviour in potential guide dogs. Appl Anim Behav Sci 15: 247–260.
View Article • Google Scholar

28. Sinn DL, Gosling SD, Hilliard S (2010) Personality and performance in military working dogs: Reliability and predictive validity of behavioral tests. Appl Anim Behav Sci 127: 51–65.
View Article • Google Scholar

29. Serpell J (1995) The domestic dog: its evolution, behaviour and interactions with people. Cambridge University Press, Cambridge. 284 p.

30. Miklósi A (2009) Dog Behaviour, Evolution, and Cognition. Oxford: Oxford University Press. 304 p.

31. McCrae RR, Costa PT (2008) The five-factor theory of personality. In: John OP, Robins RW, Pervin LA (eds) Handbook of personality. Guildford Press, New York, 159–181.

32. Draper TW (1995) Canine analogs of human personality factors. J of Gen Psychol 122: 241–252.
View Article • Google Scholar

33. Gosling SD, John OP (1999) Personality dimensions in nonhuman animals: A cross-species review. Curr Dir Psycholog Sci 8: 69–75.
View Article • Google Scholar

34. Svartberg K, Forkman B (2002) Personality traits in the domestic dog (Canis familiaris). Appl Anim Behav Sci 79: 133–156.
View Article • Google Scholar

35. Serpell JA, Hsu Y (2001) Development and validation of a novel method for evaluating behavior and temperament in guide dogs. Appl Anim Behav Sci 72: 347–364.
View Article • Google Scholar

36. Segurson SA, Serpell JA, Hart BL (2005) Evaluation of a behavioral assessment questionnaire for use in the characterization of behavioral problems of dogs relinquished to animal shelters. J Am Vet Med Assoc 227: 1755–1761.
View Article • Google Scholar

37. van den Berg L, Schilder MBH, de Vries H, Leegwater PAG, van Oost BA (2006) Phenotyping of aggressive behavior in Golden retriever dogs with a questionnaire. Behav Genet 36: 882–902.
View Article • Google Scholar

38. De Meester RH, De Bacquer D, Peremans K, Vermeire S, Planta DJ, et al. (2008) A preliminary study on the use of the Socially Acceptable Behavior test as a test for shyness/confidence in the temperament of dogs. J Vet Behav 3: 161–170.
View Article • Google Scholar

39. Duffy DL, Serpell JA (2008) Behavioral assessment of guide and service dogs. J Vet Behavior 3 186–188.
View Article • Google Scholar

40. Paroz C, Gebhardt-Henrich SG, Steiger A (2008) Reliability and validity of behaviour tests in Hovawart dogs. Appl Anim Behav Sci 115: 67–81.
View Article • Google Scholar

41. Wingfield JC (2003) Control of behavioural strategies for capricious environments. Anim Behav 66: 807–816.
View Article • Google Scholar

42. Bonier F, Martin PR, Moore IT, Wingfield JC (2009) Do baseline glucocorticoids predict fitness? Trends Ecol Evol 24: 634–642.
View Article • Google Scholar

43. Ketterson ED, Nolan Jr V (1992) Hormones and life histories: an integrative approach. Am Nat 140: S33–S62.
View Article • Google Scholar

44. While GM, Isaksson C, McEvoy J, Sinn DL, Komdeur, et al (2010) Repeatable intra-individual variation in plasma testosterone concentration and its sex-specific link to aggression in a social lizard. Horm Behav 58: 208–213.
View Article • Google Scholar

45. Korte SM (2001) Corticosteroids in relation to fear, anxiety and psychopathology. Neurosci Biobehav Rev 25: 117–142.
View Article • Google Scholar

46. Carere C, van Oers K (2004) Shy and bold great tits (Parus major): Body temperature and breath rate in response to handling stress, Physiol Behav. 82: 905–912.
View Article • Google Scholar

47. Bell AM, Hankison SJ, Laskowski KL (2009) The repeatability of behaviour: A meta-analysis. Anim Behav 77: 771–783.
View Article • Google Scholar

48. Roberts BW, DelVecchio WF (2000) The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. Psychol Bull 126: 3–25.
View Article • Google Scholar

49. Srivastava S, John OP, Gosling SD, Potter J (2003) Development of personality in early and middle adulthood: Set like plaster or persistent change? J Pers Soc Psychol 84: 1041–1053.
View Article • Google Scholar

50. Soto CJ, John OP, Gosling SD, Potter J (2011) Age differences in personality traits from 10 to 65: Big-Five domains and facets in a large cross-sectional sample. J Pers Soc Psychol 100: 330–348.
View Article • Google Scholar

51. Sinn DL, Gosling SD, Moltschaniwsky J (2008) Development of shy/bold behaviour in squid: context-specific phenotypes associated with developmental plasticity. Anim Behav, 75, 433–442.

52. Svartberg K (2005) A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Appl Anim Behav Sci 91: 103–128.
View Article • Google Scholar

53. DeWitt TJ, Sih A, Wilson DS (1998) Costs and limits of phenotypic plasticity. Trends Ecol Evol 13: 77–81.
View Article • Google Scholar

54. Stamps J, Groothuis TGG (2010) The development of animal personality: relevance, concepts, and perspectives. Biol Rev 85: 301–325.
View Article • Google Scholar

55. Bergmüller R, Taborsky M (2010) Animal personality due to social niche specialisation. Trends Ecol Evol 25: 504–511.
View Article • Google Scholar

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

57. Hennessy MB, Davis HN, Williams MT, Mellott C, Douglas CW (1997) Plasma cortisol levels of dogs at a county animal shelter. Physiol Behav 62: 485–490.
View Article • Google Scholar

58. Vazire S, Gosling SD, Dickey AS, Schaprio SJ (2007) Measuring personality in nonhuman animals. In: Robins RW, Fraley RC, Krueger R (eds) Handbook of Research Methods in Personality Psychology (pp. 190–206). New York: Guilford. Guilford, New York, 190–206.

59. Hennessy MB, Voith VL, Mazzei SJ, Buttram J, Miller DD, et al. (2001) Behavior and cortisol levels of dogs in a public animal shelter, and an exploration of the ability of these measures to predict problem behavior after adoption. Appl Anim Behav Sci 73: 217–233.
   View Article • Google Scholar

60. Highfill L, Hanbury D, Kristiansen R, Kuczaj S, Watson S (2010) Rating vs. coding in animal personality research. Zoo Biol 29: 509–516.
   View Article • Google Scholar

61. Epstein S (1983) Aggregation and beyond: Some basic issues on the prediction of behavior. J Pers 51: 360–392.
   View Article • Google Scholar

62. Fleeson W (2001) Toward a structure- and process-integrated view of personality: Traits as density distributions of states. J Pers Soc Psychol 80: 1011–1027.
   View Article • Google Scholar

63. Fleeson W (2004) Moving personality beyond the person-situation debate. Curr Dir Psychol Sci 13: 83–87.
   View Article • Google Scholar

64. Christensen E, Scarlett J, Campagna M, Houpt KA (2007) Aggressive behavior in adopted dogs that passed a temperament test. Appl Anim Behav Sci 106: 85–95.
   View Article • Google Scholar

65. Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56: 81–105.
   View Article • Google Scholar

66. Mizukoshi M, Kondo M Nakamura T (2008) Evaluation of the potential suitability of guide dog candidates by continuous observation during training. J Vet Behav 3: 193–198.
   View Article • Google Scholar

67. Arata S, Momozawa Y, Takeuchi Y, Mori Y (2010) Important behavioral traits for predicting guide dog qualification. J Vet Med Sci 72: 539–545.
   View Article • Google Scholar

68. Martinek Z, Lat J (1969) Long-term stability of individual differences in exploratory behaviour and rate of habituation in dogs. Physiol Bohemoslov 18: 217–225.
   View Article • Google Scholar

69. Borenstein M (2009) Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. New York: Russel Sage Foundation. 221–235.

70. Rosenthal R, DiMatteo MR (2001) Meta-analysis: Recent developments in quantitative methods for literature reviews. Annu Rev Psychol 52: 59–82.
   View Article • Google Scholar

71. Grubbs FE (1950) Sample criteria for testing outlying observations. J Am Stat Assoc, 21, 27–58.

72. Barnett V, Lewis T (1994) Outliers in statistical data analysis (3rd ed.). New York: Wiley. 604 p.

73. Duval S, Tweedie R (2000a) A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J Am Stat Assoc 95: 89–98.
   View Article • Google Scholar

74. Duval S, Tweedie R (2000b) Trim and fill: A simple funnel plotbased method of testing and adjusting for publication bias in metaanalysis. Biometrics 56: 276–284.
   View Article • Google Scholar

75. Hedges LV, Vevea JL (1998) Fixed- and random-effects models in meta-analysis. Psychol Methods 3: 486–504.
   View Article • Google Scholar

76. Hedges LV (1992) Meta-analysis. J Educ Stat 17: 279–296.
   View Article • Google Scholar

77. Field A (2003) Can meta-analysis be trusted? Psychologist 16: 642–645.
   View Article • Google Scholar

78. Hunter JE, Schmidt FL (2000) Fixed effects vs. random effects meta-analysis models: implications for cumulative knowledge in Psychology. Int J Select Assess 8: 275–292.
   View Article • Google Scholar

79. Lipsey MW, Wilson DB (2001) Practical meta-analysis. Thousand Oaks, CA: Sage. 264 p.

80. Cooper H, Hedges LV, Valentine JC (2009) Handbook of research synthesis and meta-analysis. New York: Russell Sage. 615 p.

81. Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Orlando, FL: Academic Press. 369 p.

82. Borenstein M, Hedges L, Higgins J, Rothstein H (2005) Comprehensive meta-analysis (Version 2) [Computer software]. Englewood, NJ: BioStat.

83. Stephen J (2006) Ph.D. Dissertation: Stress and personality in kennelled dogs: how to identify those individuals at risk of poor welfare in shelters: Brunel University, United Kingdom.

84. McPherson JA (1998) Ph.D Dissertation: Aetiology, characterisation, and prediction of separation-induced behaviour in the domestic dog: University of Southampton, Southhampton, UK.

85. Ley JM, McGreevy P, Bennett PC (2009) Inter-rater and test-retest reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R). Appl Anim Behav Sci 119: 85–90.
View Article • Google Scholar

86. Weiss E, Greenberg G (1997) Service dog selection tests: Effectiveness for dogs from animal shelters. Appl Anim Behav Sci 53: 297–308.
View Article • Google Scholar

87. Garamszegi LZ, Markó G, Herczeg G (2012) A metha-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. Evol Ecol 26: 1213–1235.
View Article • Google Scholar

88. Stamps J (2003) Behavioural processes affecting development: Tinbergen's fourth question comes of age. Anim Behav 66: 1–13.
View Article • Google Scholar

89. Soloman MB, Karom MC, Norvelle A, Markham CA, Erwin WD, et al. (2009) Gonadal hormones modulate the display of conditioned defeat in male Syrian hamsters. Horm Behav 56: 423–428.
View Article • Google Scholar

90. King T, Marston LC, Bennett PC (2009) Describing the ideal Australian companion dog. Appl Anim Behav Sci 120: 84–93.
View Article • Google Scholar

91. Naguib M, Flörcke C, van Oers K (2011) Effects of social conditions during early development on stress response and personality in great tits (*Parus major*). Dev Psychobiol 53: 592–600.
View Article • Google Scholar

92. Rödel HG, Monclús R (2011) Long-term consequences of early development on personality traits: a study in European rabbits. Behav Ecol 22: 1123–1130.
View Article • Google Scholar

93. Rödel HG, Meyer S (2010) Early development influences ontogeny of personality types in young laboratory rats. Dev Psychobiol 53: 601–613.
View Article • Google Scholar

94. Coren S (1998) Why we love the dogs we do. New York: Free Press. 320 p.

95. Galibert F, Quignon P, Hitte C, André C (2011) Toward understanding dog evolutionary and domestication history. CR Biol 334: 190 – 196.

96. John OP, Soto CJ (2007) The importance of being valid. In: Robins RW, Fraley RC, Krueger R (eds) Handbook of Research Methods in Personality Psychology. The Guilford Press, New York: 461–494.

97. Svobodová I, Vápeník P, Pinc L, Bartoš L (2008) Testing German shepherd puppies to assess their chances of certification. Appl Anima Behav Sci 113: 139–149.
View Article • Google Scholar

98. Turcsàn B, Kubinya E, Miklósi A (2011) Trainability and boldness differ between dog breed clusters based on conventional breed categories and genetic relatedness. Appl Anim Behav Sci 132: 61–70.
View Article • Google Scholar

99. Battaglia CL (2009) Periods of early development and the effects of stimulation and social experiences in the canine. J Vet Behav 4: 203–210.
View Article • Google Scholar

100. Pierantoni L, Albertini M, Pirrone F (2011) Prevalence of owner-reported behaviours in dogs separated from the litter at two different ages. Vet Rec: 1–6.

101. Koda N (2001) Development of play behavior between potential guide dogs for the blind and human raisers. Behav Process 53: 41–46.

View Article • Google Scholar

102. Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: A practice guide for biologists. Biol Rev 82: 591–605.
View Article • Google Scholar

103. Dingemanse NJ, Dochtermann N, Wright J (2010) A method for exploring the structure of behavioural syndrome to allow formal comparison within and between datasets. Anim Behav 79: 439–450.
View Article • Google Scholar

104. Wolf M, van Doorn GS, Leimar O, Weissing FJ (2007) Life-history trade-offs favour the evolution of animal personalities. Nature 447: 581–584.
View Article • Google Scholar