**Materials and Methods**

*Literature search and data collection*

We conducted a systematic review and meta-analysis by systematically searching the literature for empirical studies quantifying animal personality behavioural traits in males and females (as classified by Sih *et al*. 2004 & Réale *et al*. 2007). We collected 9,698 records from both ISI Web of Science and Scopus databases on 11 December 2018, using primary and secondary keyword searches (see PRISMA in Figure 1). A secondary search was used to find and include studies that measured ‘personality-like behaviours’; behaviours not explicitly defined as personalities but were measures of activity, aggression, boldness, exploration or sociality. We also conducted a forwards citation search in both databases to include every record that had cited one of five key animal personality reviews: (Dall *et al.* 2004; Sih *et al.* 2004; Réale *et al.* 2007; Schuett *et al.* 2010; Dingemanse & Wolf 2010).

After removing duplicates, 3,739 records were left for title and abstract screening by LH. A total of 942 papers were excluded at the title/abstract screening stage because they were: 1) duplicates (*n* = 130), 2) non-animal studies (*n* = 256), 3) reviews or non-empirical studies (*n* = 516), 4) were not published in English (*n* = 8), or 5) we couldn’t access the full-text record (*n* = 32). This left 2,797 papers for full-text screening. We included papers in the final dataset only if they provided raw means and some measurement of error for both males and females because raw summary statistics are required for the derivation of variance-based effect size measures and provide greater opportunities to control for sources of non-independence (Nakagawa *et al.* 2015; Noble *et al.* 2017). Studies that only report PCAs and factor loadings were excluded as their interpretation can be challenging (e.g. Tarka *et al*. 2018). Briefly, we excluded papers at the full-text stage if they met the following exclusion criteria: 1) were not personality studies, 2) studies used domestic or agricultural animals where individuals were neutered, 3) studies did not compare males and females (i.e. studies where animals were hermaphroditic, clonal or gynandromorphs, or where both sexes were analysed together, or only one sex was studied), or 4) data were unsuitable/missing (as shown in Figure 1).

In total, this provided 211 eligible studies with suitable data that could be included in our meta-analysis (see Figure 1). We extracted means, error (standard deviation, standard error) and sample sizes from text, tables, figures and supplementary data files for both males and females on all personality traits described by the study authors. We used the R package *metaDigitise* (v1.0.0, Pick *et al*. 2019) to extract summary statistics from figures where needed.

*Effect size and sampling error*

To understand how males and females differ in their mean personality-like traits and their variability we used two contrast-based effect size measures. To understand changes in mean personality between the sexes we calculated Hedges’ *g* (sample size adjusted standardised mean difference; Hedges & Olkin 1985) and its associated sampling error as follows:

,

Where is the mean, is the pooled standard deviation and N and SD are the sample size and standard deviation for males (M) and females (F), respectively. We used Hedges *g* instead of log response ratios as our chosen effect size as much of the data was not on a ratio scale and as such a response ratio could not be calculated. Effect size direction for a given behavioural trait depended on the way in which it was measured within a study. For example, boldness can often be measured as latency to flee, where a shorter time indicates a shyer individual. Yet the time to resume behaviour following a simulated predator approach, another common boldness measure, means that bolder individuals are those that resume normal behaviour more quickly. As such, we carefully examined all behavioural traits and ensured that the directionality (i.e., larger mean) was always consistent, flipping the direction of effect size estimates if directionality did not match.

We had strong mean-variance relationships in our data. As such, to understand differences in variance in personality traits across the sexes we used the log coefficient of variation (lnCVR) (Nakagawa *et al.* 2015), calculating the effect size and samping variance as follows:

,

where CV*M* and CV*F* are *SDM /M* and *SDF /F* , respectively. The lnCVR allows us to quantify the difference in variance between the sexes independent of changes in the mean (Nakagawa *et al.* 2015).

*Moderator variables*

We collected a series of moderator variables that we *a priori* expected to moderate the effect size magnitude and direction. Personality traits extracted for males and females were categorised into one of five types suggested by Réale *et al.* (2007); ‘boldness’, ‘aggression’, ‘activity’, ‘sociality’ and ‘exploration’. We also recorded the taxonomic group (‘invertebrates’, ‘fish’, ‘amphibians’, ‘reptiles’, ‘birds’, ‘mammals’), the age of the sample (‘juveniles’ or ‘adults’), whether the study population was from the lab or the wild, and whether the study was conducted in the lab or field given this is expected to affect behaviour (Tarka *et al*. 2018). We also generated three moderator variables that attempt to capture the strength of sexual selection. First, we quantified the degree of sexual size dimorphism (SSD) between males and females as this is strongly correlated with the strength of sexual selection (e.g. birds: Székely *et al*. 2004; Dale *et al*. 2007; fish: Walker & McCormick 2009; mammals: Promislow 1992). We created an SSD index by taking the ratio of male and female mean body size, mass or another sexually dimorphic trait (e.g. wing length), and used the following calculation (reference):

Our second measure attempting to capture the strength of sexual selection was the type of mating system (‘polygynous’, polygamous’, ‘monogamous’, polyandrous’). While mating system is a more indirect measure, a greater amount of data was available to categorise species compared with measures of direct SSD between the sexes. Lastly, we also categorised the parental care strategies exhibited by each species within and across studies (‘biparental’, ‘maternal’, ‘paternal’, ‘none’). Where body size measures for males and females, mating system and parental care were not reported in the research article itself, we obtained these data by searching Web of Science, Scopus and Google Scholar. We used the search terms: “species name” AND male AND female AND body size OR length (for body size measures), “species name” AND parental care OR mating system for parental care and mating system.

*Meta-analyses*

We modelled SMD and lnCVR using multi-level meta-analytic (MLMA) and multi-level meta-regression (MLMR) models in R using the package *metafor* (version Viechtbauer, 2010; R Core Team, 2019). In all our models we subset the data into broad taxonomic groups (i.e., birds, mammals, fish, invertebrates and reptilia (amphibians and reptiles combined due to low sample sizes)) as we were: 1) specifically interested in the extent to which patterns varied across these taxonomic groups and 2) because constructing phylogenies were easier within these sub-groups. We first estimated the overall evidence for a sex-specific effect in mean and variance across all personality traits for each specific taxonomic group using MLMA models. We accounted for effect size measurement error in our model and included study and phylogenetic random effects. We also estimated an observational-level (residual) variance as *metafor* does not estimate this by default. Phylogenetic correlation matrices were derived for each group either from existing phylogenetic tree databases (e.g. BirdTree.org for birds; Jetz *et al.* 2012, 2014) or using TimeTree.org (Hinchcliff, 2015). For the bird phylogeny, we used the Ericson tree backbone (Ericson *et al*. 2006) to generate 1,000 trees and took a sample of 100 trees. We used *ape* and *phytools* in R to take an average tree from these 100 which we used for analyses. TimeTree phylogenies were derived by exporting a list of taxa from (TimeTree.org). In some instances a species that was present in our dataset had no phylogenetic representation so we found and substituted the closest sister species. In all cases, we resolved synonymous taxa across our dataset and trees and pruned our trees where needed. From our MLMA models we derived heterogeneity estimates (*I*2; Thompson & Higgins; Nakagawa & Santos 2012). More specifically we derived between-study heterogeneity estimates (*I*2study), phylogenetic heterogeneity or phylogenetic heritability (*I*2phylo) (Nakagawa & Santos, 2012; Hadfield & Nakagawa, 2010) along with total heterogeneity (*I*2Tot), which is simply the proportion of total variance in effect size estimates excluding the total sampling variance.

We also fitted a series of models that included key moderator variables we predicted would explain variation in effect size. In all models, we again controlled for measurement error, study and phylogeny. Our first MLMR model estimated overall-meta-analytic mean estimates for each personality trait type separately (i.e. ‘activity’, ‘aggression’, ‘boldness’, ‘sociability’, ‘exploration’) within each taxonomic group. We expected that any mean and variance differences between the sexes would depend on the type of personality trait measured, especially where trait types are correlated with different sex roles (e.g. parental care; Schuett *et al*. 2010), life-histories (e.g. sex-biased dispersal and exploration or activity trait types; Cote *et al*. 2010), or are under sexual selection (e.g. male aggression might reflect the level of male-male competition or female mate choice; Munson *et al.* 2020). We also tested whether the degree of sexual selection, as measured by 1) the degree of sexual size-dimorphism (SSD) or 2) mating system, moderated effect sizes. We predicted that species with strong SSD or that have polygynous mating systems would show stronger differences in mean and variance. However, we also expected such effects to depend on the personality trait being measured. As such, we fitted a model that included personality trait type, SSD and their interaction and a second model that included personality trait type, mating system and their interaction. We present meta-analytic mean estimates and 95% confidence intervals from our models throughout.

Studies included in our meta-analysis varied greatly in their design and there was a number of additional sources of non-independence within studies (Noble *et al.* 2017). First, multiple personality traits were quantified on the same sets of individuals (e.g. ‘boldness’ and ‘aggression’). Additionally, some studies measured the same individuals multiple times for the same trait (i.e. repeatability), or for the same trait type using a different test (e.g. several measures of boldness). We conducted a series of sensitivity analyses to ascertain the impact that these sources of non-independence had on our final results. We created correlation matrices among effect sizes that shared the same sets of individuals in the sample used to derive effect sizes. Given that we did not know the exact correlation among traits we created three different dependency matrices (**D**; i.e., correlation matrices) that assumed r = 0.3, r = 0.5 and r = 0.8. We refit our models replacing the identify matrix (**I**) that was assumed when estimating the residual error variance with our **D** matrices.



**Figure 1.** *PRISMA* diagram showing the process of finding, screening and including/excluding studies for this meta-analysis. Searches were further refined to agriculture & biology (Scopus), zoology, ecology, biology, multidisciplinary sciences, evolutionary biology (WoS) categories.