**Supplementary Material for Harrison *et al*. 2020**

1. *Keyword search terms*

**Table S25.** Primary and secondary keyword searches used for our literature search conducted on 11 December 2018. Keyword searches differed slightly to account for how operator terms are employed by each database. The searches were further refined by using relevant biological categories in ISI or SCOPUS.

|  |  |  |
| --- | --- | --- |
|  | ***ISI Web of Science*** | ***SCOPUS*** |
| *Primary keyword search terms* | (personalit\* OR “behavioural syndrome\*” OR “behavioral syndrome\*” OR temperament) AND (sex\*) NOT (man OR men OR woman OR women OR human) | personalit\* OR “behavioural syndrome” OR “behavioral syndrome” OR temperament AND sex AND NOT man AND NOT woman AND NOT human |
| *Secondary keyword search terms* | animal\* AND behav\* AND (bold\* OR shy\* OR neoph\* OR aggress\* OR explor\* OR hid\*) AND “sex differences” NOT man NOT woman NOT human | animal\* AND (bold\* OR shy\* OR neoph\* OR aggress\* OR explor\* OR hid\*) AND “sex differences” AND NOT man AND NOT woman AND NOT human |
| *Refined by categories* | zoology, ecology, biology, multidisciplinary sciences, evolutionary biology | agriculture & biology |

1. *Data collection – mating system and parental care moderator terms*

We searched *ISI Web of Science, Scopus* and *Google Scholar* using the search terms: “species name” AND “mating system” for mating system and “species name” AND “parental care” for parental care. For birds, we also searched the online reference database *Birds of the World* (birdsoftheworld.org; accessed via an ANU library subscription in 2019) by searching “species name”. We noted whether the mating system of the species was characterised by ‘multiple mating’ or ‘monogamy’, and whether the species provided ‘maternal’, ‘paternal’, ‘biparental’, ‘cooperative’ or ‘no care’. However, after data collection we decided to drop parental care from subsequent analysis because 1) we did not have enough data for enough species to run our proposed meta-regression models and 2) data quality was questionable. The location of data collected for both parental care and mating system (and body sizes for SSD) are provided in the accompanying Supplementary data file ‘sexual\_selection.xlsx’.

1. *Excluding studies and effect sizes*

After full-text screening, we were left with a total of n=247 studies. However, n=29 studies were excluded from this initial inclusion list because behaviours did not really fit into personality categorisation, or were missing some key data required to calculate effect sizes. Another n=8 studies were excluded before analysis was conducted because studies were missing data required to calculate effect sizes, producing NAs. This left us with a total of n=210 studies in our final dataset.

Before analysis, we removed n=2 effect sizes (both from the same study on invertebrates) that were extremely large outliers. We also had n=24 effect sizes in our dataset that were more physiological than behavioural (i.e. breathing rate, max. heart rate after capture etc.). We decided to remove these effect sizes before running models, which reduced the total number of effect sizes, but did not change the number of studies or species in our final dataset.

While we decided not to email authors for missing data (i.e. sample sizes, error type, additional information), we did obtain sample sizes for males and females in study P077 via email correspondence with a co-author of the study.

Additionally, there were 3 studies in our dataset (P172, P210 and P231) that we were concerned might have issues with data duplication/unreliability. LMH checked the retraction database *retractiondatabase.org* regularly during data analysis to check if these studies had been either retracted or flagged for concern. As of 31 August 2020, none of the 3 studies had been retracted, nor had any other issues, so we decided to keep these studies in our final dataset.

1. *Score data*

We performed data transformations on latency data and proportional data in order to meet assumptions about normality. However, we were unable to adjust score data, and therefore assumed that scores were normally distributed. Scores did not make up a large proportion of our effect size dataset, but we decided to run contrast meta-analysis models to check whether data composed of scores were significantly different from the rest of our dataset. Summaries of these models are shown in Supplementary Table S15.

*Score data - invertebrates*

Effect sizes calculated from scores were significantly different from the rest of the dataset (including transformed latency and proportion data) for mean differences in personality for invertebrates only (intercept: =0.30, 95% CIs: 0.02, 0.57, *t* = 2.12, *p*=0.03; score: = -0.29, 95% CIs: -0.57, -0.02, *t* = -2.11, *p*=0.04). Invertebrates had the most score data of any taxonomic group (n=61 effect sizes). We therefore decided to exclude these scores from our invertebrate dataset and rerun our models. All invertebrate model tables and figures show data after scores were removed.

1. *Calculating I2*

We extracted *I*2 from our meta-analytic intercept-only models (see Supplementary Table S1) using the following equation:

Where is the total variance, is the phylogenetic variance, is the between-study variance, is the species-specific variance, is the study-specific variance (observation-level random effect), and is the remaining within-study sampling variance (random effects) (Nakagawa & Santos 2012).

Further, we can then partition *I*2 to calculate study-level *I*2 and species-level *I*2 (*I*2s and *I*2u, respectively) (Nakagawa & Santos 2012):

*I*2s = / ,

*I*2u = /

Finally, we can determine the strength of phylogenetic variance using the equation:

Where = 0 there is no phylogenetic signal, but when = 1 there is a strong effect of phylogeny on heterogeneity (Nakagawa & Santos 2012).

1. *Exploratory analysis*

We extracted information on factors that differed among studies where we had an *a priori* expectation that they might moderate the magnitude and/or direction of the effect size. For mating system, we followed methods similar to those used to obtain SSD measures. Where mating system was not reported in the research article, we searched *Web of Science*, *Scopus* and *Google Scholar* using the search terms: “species name” AND “mating system”. For birds, we first searched the online reference database *Birds of the World* (birdsoftheworld.org; accessed via an ANU library subscription in 2019). Initially we wanted to categorise mating system into the following: “monogamous”, “polyandrous”, “polygynous”, or “promiscuous”. However, it was difficult to find studies that agreed on mating system definitions for many species, so we collapsed our mating system categories into just “monogamous” or “multiple mating”. The location of data collected for mating system is provided in the Supplementary Material data files (see sexual\_selection.xlsx).

1. *Exploratory analysis - results*

*Mating System*

Monogamous and multiple mating systems were not significantly different from each other for means or variability for any of the taxonomic groups (Supplementary Table S9).

*Age*

Adults and juveniles were not significantly different from each other for means or variability for any of the taxonomic groups (Supplementary Table S10).

*Population*

Fish from the wild had greater differences in variability than fish from lab populations (intercept: =-0.09, 95% CIs: -0.18, -0.01, *t* = -2.11, *p*=0.04; lab: =0.08, 95% CIs: -0.02, 0.19, *t* = 1.55, *p*=0.12), but not for mean personality differences, and not for any other taxonomic group (Supplementary Table S11).

*Study environment*

Studies conducted in the lab were significantly different to field studies for mammals (intercept: =0.24, 95% CIs: -0.10, 0.57, *t* = 1.38, *p*=0.17; lab: = -0.30, 95% CIs: -0.56, -0.04, *t* = -2.26, *p*=0.02), but not for variability, and not for any other taxonomic group (Supplementary Table S12).

*Study type*

Effect sizes from observational studies were significantly different from experimental studies for mammals (intercept: =0.00, 95% CIs: -0.21, 0.22, *t* = 0.04, *p*=0.97; observation: = 0.38, 95% CIs: 0.12, 0.64, *t* = 2.81, *p*=0.005), but not for variability, and not for any other taxonomic group for which comparisons could be made (Supplementary Table S13).

1. **D** *matrices – results*

*Intercept only models*

Regardless of whether the correlation between personality traits was set to either rho=0.3, 0.5 or 0.8 (i.e. to control for traits that were measured on the same individuals within the same study), there remained no significant sex difference between either the mean or the variability for personality overall, for any of the five taxonomic groups (see Supplementary Tables S16-S18). We therefore interpreted our initial intercept only models without **D** matrices (Supplementary Table S1).

*Personality trait models - mean differences*

Female birds were more social than males, while male reptiles/amphibians were more explorative than females (at each of the levels of rho; see Supplementary Tables S19-S21). These significant sex differences, and their direction, were consistent with those from models without any **D** matrix (Supplementary Table S2). However, male invertebrates were more active than females when rho=0.3 (Supplementary Table S19), when rho=0.5 (Supplementary Table S20), and when rho=0.8 (see Supplementary Table S21). This significant sex difference, for invertebrates, was not found in our initial MLMR personality models (but was marginally significant, see Supplementary Table S2).

*Personality trait models - variability*

Adjusting the levels of rho for variability did not significantly change the results obtained from initial MLMR personality models for any of the taxonomic groups. Importantly, female fish remained significantly more variable than males for aggressive behaviour, regardless of the level of rho (see Supplementary Tables S19-S21).

*SSD and personality - should probably run these on our reduced models*

Ran these D matrix models on the full models with interactions and nothing changed. Should I do the subsetted models instead?

1. *References*

Nakagawa, S. & Santos, E.S.A. (2012). Methodological issues and advances in biological meta-analysis. *Evol. Ecol.*, 26, 1253–1274.