**Results**

*The effect size dataset*

Looking at the format for Nat Comm (in Cally’s paper), this first paragraph should outline/breakdown the dataset – what kinds of numbers do we have overall? Per personality trait? Per taxo group? What did we get rid of? Lots of n’s in here… Mention that because our dataset covers a broad range of taxonomic groups, and because these animal groups are all so different (i.e. ectotherms and endotherms, invertebrates and vertebrates, heterogametic males and homogametic males) we thought it best to split the dataset by taxonomic group for all analysis.

*Hypothesis 1/Model 1 – Personalities do not differ between the sexes across all taxonomic groups*

The overall means across all personality traits were not significantly different from zero in any of the taxonomic groups (restricted maximum likelihood (REML) birds: *β* = -0.09, 95% CIs: -0.33, 0.14; fish: *β* = -0.02, 95% CIs: -0.38, 0.33; invertebrates: *β* = 0.20, 95% CIs: -0.08, 0.48; mammals: *β* = 0.09, 95% CIs: -0.25, 0.43; reptiles: *β* =0.06, 95% CIs: -0.11, 0.22), indicating an absence of sex differences in mean personality behaviours.

The overall mean differences in variability, across all personality traits and for all taxonomic groups, were also non-significantly different from zero (restricted maximum likelihood (REML) birds: *β* = -0.16, 95% CIs: -0.67, 0.34; fish: *β* = -0.02, 95% CIs: -0.08, 0.04; invertebrates: *β* = -0.04, 95% CIs: -0.12, 0.04; mammals: *β* = 0.07, 95% CIs: -0.20, 0.33; reptiles: *β* =0.02, 95% CIs: -0.07, 0.12). We observed high heterogeneity (total I2) within each taxonomic group for both SMD and lnCVR datasets (see Table ??). For the SMD dataset, heterogeneity was mostly explained by between study variance by including Study ID in our models (see Table ??), with exception for fish; heterogeneity was attributed mostly to phylogenetic relatedness (total I2 = 72%, 95% CIs: 68-75%; phylogenetic I2 = 23%, 95% CIs: 13-35%). Interestingly, heterogeneity within the lnCVR dataset for birds and mammals was explained by phylogenetic relatedness (birds: total I2 = 94%, 95% CIs: 93-94%; phylogenetic I2 = 46%, 95% CIs: 39-54%; mammals: total I2 = 59%, 95% CIs: 54-64%; phylogenetic I2 = 41%, 95% CIs: 30-51%), while heterogeneity remained mostly unexplained for the remaining taxonomic groups (see Table ??).

*Hypothesis 2/Model 2 – Sex differences in the animal Big Five*

Model 2 – taxo group and personality trait …

*Hypothesis 3/ Model 3 – Sexual size dimorphism alone doesn’t moderate sex differences in the mean or variability of personalities, but interacts with some traits to produce sex differences.*

Model 3 – SSD and personality traits and taxo group, and their interactions

*Sensitivity analyses and robustness*

When I do them, put them in here …

*The effects of mating system, age, and other study characteristics on personality means and variability*

Secondary models were conducted to understand how mating system (monogamy or multiple mating), study animal age (juvenile or adult), study population (wild or lab/captive), study environment (field or lab) and study type (observational or experimental) might moderate sex differences in means and variability for personalities in each of the five taxonomic groupings. Some of these categories have previously been implicated to influence sex differences in mean and variability of animal personalities (e.g. Tarka *et al*. 2018), so they were also included in our analysis for confirmation. It is important to note that there was inadequate data to compare any secondary model categories at the personality trait level within taxa.

Surprisingly, neither mating system nor age explained sex differences in mean personality or the variability in any of the taxonomic groups (see Table ?? for slope estimates and 95% CIs). However, female fish from wild populations had significantly greater variability in their personality compared to males from the wild (wild: *β* = -0.10, 95% CIs: -0.19, -0.01, *p*=0.03), while lab-reared fish populations had no sex differences in variability (lab/captive: *β* = 0.03, 95% CIs: -0.05, 0.10, *p*=0.68). Such a sex difference was absent in the mean personalities of both wild and lab fish populations (wild: *β* = -0.11, 95% CIs: -0.47, 0.25; lab/captive: *β* = 0.03, 95% CIs: -0.33, 0.38), and there was no effect of population on sex differences in the mean or variability of personalities in any other taxonomic group (see Table ??). Study environment had a marginally significant effect on mean personality in mammals; males in the wild had somewhat larger overall means than females (*β* = 0.27, 95% CIs: -0.02, 0.56, *p*=0.06), a difference that was absent from mammals in a laboratory environment (*β* = -0.06, 95% CIs: -0.35, 0.23, *p*=0.66). Study environment did not influence sex differences in mean personality in any other taxonomic group, or the variability of any taxonomic group (including mammals; see Table ??). Finally, whether a study was observational (no manipulation) or experimental (any kind of experimental manipulation) explained sex differences in the mean personality of mammals; males had greater means than females in observational studies compared to mammals in experimental studies (experimental: *β* = 0.01, 95% CIs: -0.12, 0.14, *p*=0.90; observational: *β* = 0.43, 95% CIs: 0.21, 0.65, *p*<0.001). Again, there was no effect of study type on variability for mammals in either experimental or observational studies, nor for the means or variability in any other taxonomic group (see Table ??).

*Publication bias*

Funnel plots and Egger’s regression test were used to look for potential publication bias in our standardised mean (Hedge’s *g*) and variability (lnCVR) effect size datasets. The Egger’s test suggested the presence of publication bias for the standardised means (*z* = -2.79, p= 0.005), but not for the variability dataset (*z* = -0.03, p = 0.97). However, significance generated from the Egger’s test can also be an indicator of high heterogeneity, which is the case for our effect sizes. As an additional check, we looked at the correlation between effect size (Hedge’s *g*) and journal impact factor, and the correlation between effect size and publication year…