Dear Authors,

Your manuscript has been reviewed by two referees and based on their reviews, I am inviting you to revise it according to their comments and suggestions.

Notably, reviewer 2 provided important comments and guidelines to improve the reproducibility and transparency of the work.

Both reviewers also provide important suggestions to improve the readability of the manuscript. Reviewer 1 suggests to more clearly state in the abstract what the four main predictions were and whether the results confirmed these predictions or not both in the abstract and the discussion. Reviewer 2 also provide suggestions for improving the readability of the introduction.

Finally both reviewers suggests ways to better highlight the potential implications of the study, among other suggestions by providing a general conclusion in the abstract, summarising the limitations of the study, and speculate about how these findings might relate to non-mammalian social species.

You will find their reviews below (I take the opportunity to thank the reviewers!). In addition, I also have some minor comments:

Reply: *Dear Dr Paquet*

We are thankful to both reviewers and to you for these constructive and helpful comments. We responded to your comments below (replies are in italics).

We made these changes to the file at https://doi.org/10.32942/osf.io/rc8na

As before, the version-tracked file is in rmarkdown at GitHub: https://github.com/dieterlukas/FemaleDominanceReproduction_MetaAnalysis/blob/trunk/Man uscriptfiles/PostStudy_MetaAnalysis_RankSuccess.Rmd. In case you want to see the history of track changes for this document at GitHub, click the previous link and then click the "History" button on the right near the top. From there, you can scroll through our comments on what was changed for each save event and, if you want to see exactly what was changed, click on the text that describes the change and it will show you the text that was replaced (in red) next to the new text (in green).

We think the revised version is much improved due to your generous feedback!

All our best,

Shivani, Elise, and Dieter

Comment 1: Lines 42-44: statement (4); it would be useful to describe more clearly this effect here (in which direction does the effect go).

Reply 1: We had previously used the more general statement because we found that multiple variables that reflect variation in female sociality are associated with variation in the effect size. We have now attempted to summarize these variables by referring to the kinds of societies they represent:

Abstract: "Instead, we found that the social environment consistently mitigates rank differences on reproductive success by modulating female competition, with, as predicted, (4) dominant females showing higher reproductive success than subordinates in two different extreme types of societies: first, effect sizes are highest when societies are structurally complex which occurs when females live in cooperatively breeding groups with many individuals; second, they are also elevated when relationships among females in the societies are complex which occurs when groups are composed of unrelated females."

Comment 2: Line 524: why? Please justify this change.

Reply 2: Before we had run the analyses, a collaborator on a different project alerted us to the discussions around how to calculate sexual dimorphism. We now explain this:

Changes from preregistration > additional variables: "We changed how we calculated sexual dimorphism in body weight. We had previously taken the ratio of male weight divided by female weight. A collaborator on a different project, in which we also use sexual dimorphism in body weight as a variable, alerted us to the article by @smith1999statistics which shows that this simple ratio is biased because its distribution across species is non-linear resulting in asymmetries when females are the larger sex (as example, assume a species where individuals of one sex are 10kg and individuals of the other sex are 8kg; if males are the larger sex the simple ratio would indicate that the larger sex is 25% larger [10/8=1.25]; however, if females were the larger sex it would indicate that the larger sex is only 20% larger [8/10=0.80]). We therefore switched to the calculation of sex ratio as recommended in this article by calculating sexual dimorphism as the average weight of males divided by average weight of females if males are heavier than females and as two minus the average weight of females divided by the average weight of males otherwise."

Comment 3: Lines 580-582: it is not clear to me whether these tests were performed here for the current study, or whether the authors refer to the outcome of the cited reference.

Reply 3: We have changed the text to make it clearer that these are outcomes of analyses we performed as part of the current study:

Results > Sample bias: "We applied tests for 'publication bias' that expect a standard distribution of p-values (@preston2004adjusting) to our data"

2

Comment 4: Line 586 and elsewhere in the manuscript: what are these values? mean plus minus sd? confidence/credible intervals? If the latter, is it 95% confidence AND credible intervals (in rethinking, the default may be 89%Cri)? Just indicate this for the first case, so that we know for the rest of the result section.

Reply 4: Yes, they are the confidence/compatibility intervals. We did not explain this in the previous version. We now added information about these intervals to both the methods and the results:

Methods > Analysis: "We determined whether a variable had a relationship with the variation in the effect of dominance rank on reproductive success when the interval (for metafor the 95% confidence interval of the estimate; for rethinking the 89% compatibility estimate of the posterior sample) of the estimated association did not cross zero"

Results > Sample bias: "As a further indication of 'publication bias', we find that studies with small sample sizes and small effect sizes (those that presumably did not reach statistical significance) are missing in our dataset such that the average effect sizes at smaller sample sizes are more extreme than those at larger sample sizes (metafor estimate 95% confidence interval lower -0.03 to upper -0.02, rethinking estimate 89% compatibility estimate of posterior sample lower -0.09 to upper -0.04)"

Comment 5: Lines 619-621: "there is a strong effect". Are these "strong" effect sizes? Classically (e.g. according to Cohen, or Möller & Jennions) those would be referred to as small-medium effect sizes.

If the authors are referring to the statistical "clarity" (strength of evidence) of the effect rather than the size of the effect, then I suggest defining these effects as e.g. "clear effects" or "strong evidence" rather than "strong effects".

Reply 5: Thank you, yes, we meant to refer to the strength of evidence rather than the strength of the signal itself. We have changed this to say:

Results > Overall effect: "Across our sample, there is consistent evidence that females with higher dominance rank have higher reproductive success (metafor estimate of overall effect size"

Comment 6: Line 622: it is not clear to me why it is not the case for the Bayesian model and why this would cause a bias in this direction. How did the authors identify the cause of this difference in effect size?

Reply 6: Our interpretation for the different estimates from the models is based on how these models are constructed. The metafor approach is based on the distribution of the effect sizes essentially following a normal distribution (which is symmetric), with deviation from this taken as indication of sampling bias and influence of the predictor variables. The rethinking approach does not have an expected distribution for the effect sizes but instead pools information across them. The analyses for sampling bias we performed and present first

show that our sample is clearly biased, such that in models in which we include no additional predictors the two approaches are expected to give different estimates. We added further explanations:

Results > Overall effect: "(the metafor estimate here and in the additional models is lower than the rethinking estimate because the statistical approach of the former expects the data to be more symmetrical than they are (see Figure 1 for the bias) while the rethinking approach pools information from the available data, with the metafor estimate being closer to the median of the raw data of 0.23 and the rethinking estimate closer to the mean of 0.29)"

Comment 7: Also, were any model fit assessment performed (e.g. posterior predictive checks for the Bayesian models)? It is important to evaluate models' fit and report the outcomes.

Reply 7: We had tested our models during the preregistration on a simulated dataset to check that the models are specified correctly to recreate the expected patterns in the data. We now explain this more clearly in the manuscript:

Methods > Analysis: "the github repository also contains a simulated dataset with the same structure as the actual data, which we used to assess the fit of our models in the preregistration."

We had also compared implied predictions to the raw data, seeing that the median values we calculated matched the estimates and that estimated slopes fit to the raw data (e.g. figure 9). In light of this comment, we now added a posterior predictive check, similar to the predictive intervals from metafor displayed in the orchard plots, for the model with the clearest difference (comparison of effect sizes across breeding systems) - the result is displayed below (red dots are observed effect sizes, black lines are prediction intervals from the matching Bayesian model):
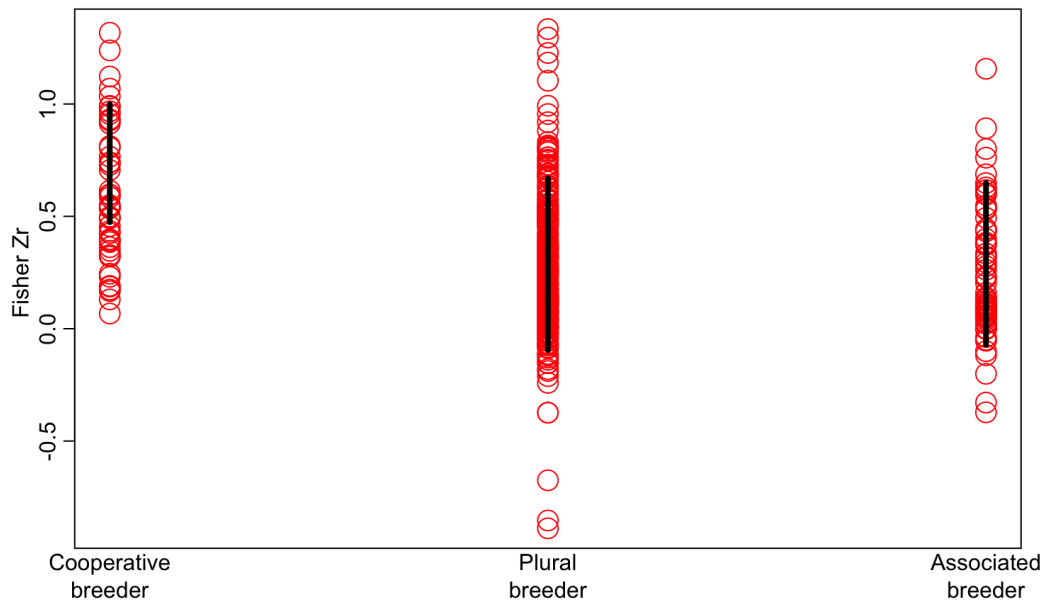
Comment 8: Lines 630-631: are these effect sizes provided somewhere? I couldn't find them but if they are, please refer to them, and if not please provide them.

Reply 8: They are the overall effect size presented in the previous section. We have now changed the text to make it clearer what is being compared to what:

Results > Influence of locality/species: "The estimate of the overall effect size did not change in this model accounting for non-independence (metafor estimate of overall effect size when accounting for non-independence +0.22 - +0.31, rethinking estimate +0.26 - +0.35) from the overall effect estimated in the base model."

Comment 9: Lines 691-692: one can only say this when looking at estimates of the difference (between the effect on survival and the effects on other measures) and the confidence/credible interval of this difference, as the authors nicely did later on in the manuscript. So please either do this here too or rephrase/delete.

Reply 9: Yes, thank you, we have now added the information about the contrasts between the different measures:

Results > R2.1 Influence of measure of reproductive success: "Effects of dominance rank on survival are lower than on other measures of female fitness (contrasts between infant survival and age at first conception/life time reproductive success/interbirth interval/infant production do not cross zero; contrasts between adult survival and age at first conception/life time reproductive success/interbirth interval do not cross zero)."

5

Comment 10: Line 692: is it meant that the effect is higher on adult than on infant survival? it reads as if adult survival is higher than infant survival. which does not mean that the effect of dominance rank is higher on adult survival. Perhaps simply remove this statement in the brackets? Also see my previous comment. given the overlapping confidence intervals of these two estimates, I doubt that there is evidence for a difference.

Reply 10: Yes, this was unclear and the estimates overlap. We have removed the sentence "In addition, females themselves appear to benefit more than their offspring (adult survival > infant survival)."

Comment 11: Line 722: The difference seems quite high given that the mean effect size of carnivores and omnivores differ so little (0.07). Can you confirm this is correct?

Reply 11: Thank you for noticing this. We had indeed miscalculated these contrasts. We have recalculated this and changed it to:

Results > R3.1 Diet Category: "Effect sizes are larger in carnivores (0.35; n=72 effect sizes) than in omnivores (0.28; n=227 effect sizes), herbivores (0.25; n=117 effect sizes), or frugivores (0.21; n=28 effect sizes) (estimated difference carnivores versus omnivores rethinking -0.14 - -0.01, difference carnivores versus herbivores rethinking -0.16 - -0.03, difference carnivores versus frugivores rethinking -0.24 - -0.02; estimates for all other comparisons cross 0)."

Comment 12: Line 729: no "clear/evidence for" associations between… One shall not conclude there is no difference (equivalent to accepting the null hypothesis).

Reply 12: We agree, which is why we had written that "our data shows no association", not that there is no association. To make it clearer, we have now changed this to:

Results > 3.2 Environmental harshness: "We found no association between environmental harshness and the effect of dominance rank on reproductive success"

Comment 13: Line 798: no differences "detected" (similar issue as previous comment).

Reply 13: We have changed this to

Results > R4.7 Coalition formation: "with no difference detected in models accounting for similarity due to phylogenetic relatedness"

Comment 14: Line 847: same again "not clearly associated". Also write in the legends or tables that it is the 95% compatibility estimate (not e.g. 89% or any other possible threshold).

Reply 14: We change the legend for the now combined table (see reply 22) to:

Table 1 legend: "Overview of results of in univariate analyses indicating whether **we did or did not find an association between individual variables with variation in effect sizes of dominance rank on female reproductive success**. The table presents for each variable which direction of association we predicted, the association we observed, and the respective estimates of the association when accounting for shared phylogenetic history among the species in our sample."

and changed the headers in the table to:

"Metafor 95% confidence interval of estimate of association" and "Rethinking 89% posterior compatibility estimate of association"

Reviews

Reviewed by anonymous reviewer, 12 Nov 2021 16:14

The preprint "The effect of dominance rank on female reproductive success in social mammals" is a very interesting meta-analysis that investigates whether high ranking social female mammals have a higher reproductive success than low ranking females. The methods and predictions outlined in the preprint have been preregistered and the authors closely follow their initially proposed methods and predictions. The authors only made minor changes to the preregistered study, including adding additional explanatory variables, a change in the presentation of the result and a few additional post-hoc analysis after confirming the predicted effect that rank effect on reproduction is more pronounced in cooperative breeding species.

I really enjoyed reading the manuscript and I am particularly happy that the authors closely followed their initially preregistered study design and study methods. As I have already reviewed their preregistration and the authors considered all of my previous comments regarding their predictions, I have only minor comments regarding the presentation and interpretation of the results.

Thank you for your feedback both on the preregistration and on this manuscript.

Comment 15: (1) Abstract: I think the abstract could be strengthened if the authors add whether these four presented results confirmed their initial predictions or not. Furthermore, I would suggest adding that the study is based on 4 main predictions and how many were confirmed or not. Additionally, I would suggest adding a general conclusion at the end of the abstract. Maybe just one or two sentences highlighting the relevance of the results and how these results help us to better understand the evolution of sociality (social ranks) in mammals?

Reply 15: We now clarify for each of the four predictions whether the observed patterns confirms or contradicts our initial expectations:

Abstract: "As predicted, we found that (1) dominance rank is generally positively associated with reproductive success, independent of the approach different studies have taken to answer this question; and that (2) life-history mechanisms mediate the relationship between rank and reproductive success, with higher effects of dominance rank on reproductive output than on survival, particularly in species with high reproductive investment. Contrary to our prediction, (3) the fitness benefits to high-ranking females appear consistent across ecological conditions rather than being elevated when resources are expected to be limited. We also found that the social environment consistently mitigates rank differences on reproductive success by modulating female competition, with, as predicted, (4) dominant females showing higher reproductive success than subordinates in two different types of complex societies: first, effect sizes are highest when societies are structurally complex which occurs when females live in cooperatively breeding groups with many individuals; second, they are also elevated when relationships among females in the societies are complex which occurs when groups are composed of unrelated females."

In addition, we added a concluding statement to the end of the abstract:

Abstract: "Our findings indicate that obtaining a high ranking position in a social group consistently provides female mammals with fitness benefits, even though future studies might show lower effects given which might have been overestimated in the literature due to various biases in the literature we were able to access, including, but not restricted to, a publication bias. They further draw a complex landscape of the level of social inequality across mammalian societies, reflected by variation in the benefits of social dominance, which appears to be shaped by reproductive and social competition more than by ecological competition."

(2) Introduction:

I am aware that the introduction is the same as in the preregistered report and this is great. While reading it again I would like to make two minor suggestions which I might have overlooked previously when I was more focused on the predictions and methods.

Comment 16: L55-L58: The meaning of the second part of this sentence is unclear to me and in particular how this relates to the predictions.

Reply 16: The sentence was a bit unclear. We have reformulated to attempt to explain better the question that motivates the predictions, and linked it directly to the following paragraph that provides some more details on the question:

Background: "However, it has remained unclear whether and when there is selection on all females to compete for a high rank. Females differ in their intrinsic qualities as well as in their access to social opportunities, which influence their ability to compete against other females. Accordingly, instead of direct selection on females to compete over high dominance rank because it provides substantial fitness benefits, selection might be on females to find

their place in the hierarchy that provides them the highest fitness given their relative characteristics."

Comment 17: L57-77: Maybe a clearer statement on how this particular study differs from any previous one (e.g. Majolo 2012) would help to understand the "novelty" of the study. Particularly the meaning of this last sentence is a bit unclear to me and how this statement relates to the previous studies.

Reply 17: We have added two sentences to the end of the paragraph to clarify that our study looks across all social mammals, not just primates, and that we include variables that were not part of the previous study:

Background: "Here, we investigate the extent and sources of variation in the effect of dominance rank on female reproductive success across all social mammals. Our study builds on the long history of research on dominance interactions (@strauss2022domarchive) by bringings together effect sizes of the relationship between rank and reproductive success from diverse mammalian societies, and we add socio-ecological predictor variables that have not been included in earlier analyses."

(3) Objectives:

Comment 18: The predictions and objective should now be in past tense.

Reply 18: We changed the phrasing in the predictions and objectives to past tense.

(4) Changes from the preregistration

Comment 19: L517-523: It would be good if the authors would explain a little bit more about the rational to include a separate analysis on macaques and how this relates to their other general predictions.

Reply 19: Following a comment of Reviewer 2 (see comment 47), we now connect the specific predictions directly with their results. As part of this, we have added a section to the results of the macaque dominance styles.

Results > Prediction 4.3: "We added an analysis after the preregistration, focusing on variation in dominance style among macaques. Macaque species have been assigned to a four-grade social style according to the relationships among females. Grade 1 species, the most despotic, are characterized by steep dominance hierarchies and more asymmetries in social interactions among breeding females, whereas grade 4 species show more frequent counter-aggression from subordinates towards dominants and less bias in social interactions. We expected that the steeper hierarchies in more despotic species would lead to larger differences in access to resources, and accordingly higher reproductive success for dominant females."

(4) Results:

Comment 20: L586-589: It's not clear to me how the authors came to this conclusion by adding sample size to the model. I understand that the positive effect size is independent of sample size (because controlled for in the model) but how can one conclude that if more effect sizes are added (particularly effect sizes that are low and that are derived from small sample sizes) would not change the relationship?

Reply 20: You are correct that additional effect sizes could change the estimate of the overall effect. We were focusing on the studies that are seemingly missing from our dataset due to publication bias. We have clarified this to say that the observed effect sizes of dominance rank on female reproductive with large sample sizes are consistently larger than zero. This suggests that the overall effect size would remain positive, even if those small or negative effect sizes that appear to be missing from the published literature were integrated. This can also be seen in Figure 2, where the line of the estimate does not cross the intercept. We have rephrased this paragraph.

Results > Sample bias: "Studies with smaller sample sizes have a higher risk to report inflated effect sizes due to a higher likelihood of Type I and Type II errors. In our dataset, the average effect sizes at smaller sample sizes are more extreme than those at larger sample sizes (effect sizes range from -0.89 to +1.33 for studies with a sample size of 20 or smaller, while for studies with sample sizes larger than 20 they range from -0.37 to +1.24). However, it is not just that the spread of values is larger for studies with smaller sample sizes, but the positive bias in effect sizes we observe decreases with the sample size of studies (metafor estimate 95% confidence interval lower -0.03 to upper -0.02, rethinking estimate 89% compatibility estimate of posterior sample lower -0.09 to upper -0.04) (Figure 2). This supports a 'publication bias', where studies with small sample sizes that did not show a positive effect are missing from the literature. However, the estimate of the intercept and slope of this model linking effect size to sample size shows that across the range of sample sizes the estimate of the overall effect size does not go below zero (see line in Figure 2). This indicates that, if studies with larger sample sizes accurately reflect the actual effect, females with higher rank have higher reproductive success."

Comment 21: Figure 5: Please indicate in the figure or in the legend what exactly the dots (means) including the thick and narrow lines represent.

Reply 21: We have added an explanation to the legend of Figure 5 to explain what is presented:

Legend Figure 5: "As in previous figures, each dot represents a single effect size, with the size of the dot indicating the precision (legend bottom right). For each measure of reproductive success, the darker circle in the middle represents the estimated mean effect, with the bold lines representing the confidence interval of the mean effect and the thinner lines the prediction estimate of the model."

Comment 22: Table 1: I really think the table is very helpful and needed. I would also suggest to include (or divide) in the table whether the effect was predicted or not (mainly for Table 1) but also the predicted effect size could also be included in Table 2.

Reply 22: Based on this helpful comment, we decided to only include a single table (Table 1) with separate rows for each variable that shows which prediction they link to, the predicted and observed effects, as well as the output from the statistical analyses. We agree that this additional information can help to provide the reader with a quicker assessment of the overall outcomes of the univariate analyses.

Comment 23: Figure 9: When looking at Figure 9, I was wondering how much influence the two points on the far right for plural breeders and for cooperative breeder have on this relationship.

Reply 23: The two effect sizes for cooperative breeders with large group sizes are from the same study of naked mole-rats. In our preregistration, we stated that we would not exclude any data point, which is why we present the results as such in the manuscript. In light of your comment we now repeated these analyses without these two effect sizes. They show the same outcome as the results we report, except that the confidence interval for the slope of the relationship between group size and effect sizes in cooperative breeders now crosses zero (but lies mostly above zero, still indicating larger effect sizes when group sizes are larger).

(5) Discussion

Comment 24: In general, I think the discussion and in particular the first paragraph could be strengthened by not only summarizing all the results but also summarizing which general prediction has been supported and which not. Again, highlighting that the study investigated previously formulated clear predictions.

Reply 24: Thank you for the helpful suggestion. We have changed the first paragraph of the discussion, linking it back to the preregistered objectives of our study.

Discussion: "Our results provide support for three of our four pre-registered objectives. First, we find that in social mammals, dominant females have higher reproductive success than lower-ranking females. While there appears to be publication bias in the dataset we put together, this bias is unlikely to be so strong as to negate the overall positive effect size of rank on reproductive success. Second, positive effects of dominance rank are present across all life history measures and are highest for life-time reproductive success among plural breeders, where data for all measures of reproductive success exist. This suggests that even if dominants might face some trade-offs (e.g. higher stress levels @cavigelli2003female), obtaining a high ranking position in a social group generally leads to

fitness benefits, though how females obtain these benefits (e.g. shorter interbirth intervals versus larger offspring) differs between populations. Effects are particularly pronounced in species in which females produce large numbers of offspring at once. Third, and against our predictions, we did not find that ecological factors play a major role in mediating the benefits of rank on reproduction. Fourth, the types of society females live in appear to have a particular modulating influence. Strong associations between dominance rank and reproductive success are consistently found among cooperative breeders, they are intermediate in stable groups with small numbers of unrelated breeding females, and lowest when large numbers of females associate."

Comment 25: In my opinion the discussion could also end with a broader perspective on the evolution of social systems and potentially sociality in mammals in general. Here, or elsewhere, I would furthermore suggest that the authors speculate a little bit about how these findings might relate to other non-mammalian social species.

Reply 25: While we do not want to speculate about processes in other taxonomic groups, we have added links in places where similar patterns have been described outside of mammals, e.g.

Discussion: "Similar to what has also been found in eusocial insects (@rubenstein2016discrete) and cooperatively breeding birds (@riehl2017kinship), among cooperatively breeding mammals there usually is a single breeding dominant female and large groups occur when her reproductive output is high without loss of reproductive control, whereas among plural/associated breeding mammals groups grow large as more females/matrilines join a group leading to reduced opportunities to control reproduction for dominants."

In addition, we added references to reproductive skew theory

"Our results also show that other social factors, such as the relatedness among females, have less of a role on the effect sizes in cooperative breeders than in plural breeders, in line with theoretical predictions that complete monopolization of reproduction can be stable if subordinates are queuing to inherit the dominant position themselves (@kokko1999)."

"It appears that in situations of strong nepotism females in a group might have more similar reproductive success…as predicted when offspring production is costly (@cant1999costly)."

We also added a st to have a broader perspective on these patterns in the context of the evolution of social systems.

Discussion: "For future studies, detailed long-term investigations are not only relevant to understand the long-term consequences of the effect of dominance rank on reproduction, but also to infer the multiple mechanisms that can link rank to reproductive output (e.g. @fedigan1983dominance, @pusey1997influence, @tibbetts2022establishment). Tracing

such differences in reproductive success over multiple generations will also be important to determine the selection processes shaping social evolution."

Comment 26: L1086-1088: I think there is a word missing in this sentence

Reply 26: Thank you, we have changed this:

Discussion: "While reproductive control appears important in explaining high reproductive success of dominant females, we did not find that effect sizes differed according to how females acquire and maintain rank. Effect sizes were similar when dominant females acquire their position by kin support versus aggression or age, and similar across macaque species with different dominance styles."

Comment 27: L1098-1100: Its not clear to me how this conclusion can be drawn. i.e. do the authors assume that coalitions require complex relationships and aggression is higher in smaller groups? I think, to avoid confusions by the readers, it would be good to explain the rationale behind this conclusion a bit more.

Reply 27: We realize that we had not provided sufficient information in this section. We wanted to establish a link with previous findings. One of us (DL) had in an earlier manuscript found that across social mammals there seem to be at least two types of social complexity (cooperative breeders with high structural complexity versus groups of unrelated breeding females with high relationship complexity). The sets of variables we observed here in these analyses appear to match on to those species with high relationship complexity (female dispersal, low relatedness, coalitions). We have now clarified this by first stating the sets of variables we analyzed here and how these might link to the previous findings:

Discussion: "Instead, in our sample we observe relatively strong effects of high dominance rank in breeding systems where females form social bonds with unfamiliar/unrelated individuals they encounter when joining new breeding units upon reaching maturity (e.g. @cameron2009social), such as among equids and gorillas. Groups in which females compete with and form complex bonds with unrelated females tend to be characterized by high relationship complexity (@lukas2018social). Rates of aggression tend to be high and dominance relationships are often based on age differences with rare changes in the hierarchy, such that females who obtain high ranking positions in these units are likely to gain fitness benefits for extended periods of time."

Reviewed by anonymous reviewer, 27 Nov 2021 04:18

First, thanks for the opportunity to review for PCI, and my apologies to the authors for being a little slow to write this review.

The authors present a meta-analysis on the effects of female dominant status on reproductive success in social mammal species. The size of the dataset (187 studies on 86 mammal species) is impressive and is a useful contribution to the field. I commend the authors for preregistering their predictions and analyses. I have a number of suggestions for improving the readability of the preprint, and its reproducibility. I'll start with the reproducibility.

Thank you for your feedback and in particular for your advice on how to improve the reproducibility of our manuscript.

USEABILITY OF SHARED DATA AND CODE

Comment 28: The files on the github repository are not as helpful as they could be. First, I can't find a metadata file describing the meaning of all the columns in the data files.

Reply 28: In line with the rules at PCI Ecology, we had deposited our data in a designated repository (see 'Data and Code Availability' section in our manuscript, with the DOI for the previous version of the dataset: https://doi.org/10.5063/F1RB7312). We had chosen the Knowledge Network for Biocomplexity (KNB), using their Ecological Metadata Language to provide metadata both for the overall dataset and for the specific variables. We keep copies of the data files on github, as they are easier to access from there for direct inclusion in analyses. We have updated the dataset with the raw data according to your suggestion (see reply 29), also updating the metadata accordigly, and again provide this file on KNB.

Comment 29: Second, I can't find a raw data file, containing the information extracted from the studies prior to processing (e.g., the inferential statistics that were converted into Zr), or code to do the processing of raw data. (I note it's a requirement of PCI Ecology to make raw data available in a repository with a DOI: https://ecology.peercommunityin.org/PCIEcology/help/guide_for_reviewers).

Reply 29: We had deposited all data necessary for the analyses with a DOI prior to submitting this manuscript for review at PCI Ecology. We had thought that the necessary and relevant data for our study are the Zr effect sizes (which we had already provided in the preregistration). Our study did not generate new data and rather relied on previously published information. As mentioned in the 'Methods', we used the online calculator from the Campbell Foundation for many of the conversions or calculated these manually using the formulas in the referenced literature. Accordingly, there is no code for the conversion of the statistics. In response to your comment, we have now added the additional information on the original data that were used to generate the Zr effect sizes, describing the type of statistics provided in the respective article, the type of estimate that we extracted, and the value of the estimate. For articles where we relied directly on observational data provided in the articles (e.g. authors state that X of Y offspring of dominant females and A of B offspring

of subordinate females died) rather than published inferential statistics, we have made this clear in the dataset.

Comment 30: Third, the Rmd files are not written in a way that makes it easy for someone else to come along and use them. For example, the package cowplot is not loaded despite plot_grid being used, so this will throw an error.

Reply 30: Thank you for checking the Rmd file. We have now added the necessary library. We went through the whole code to check for all such dependencies. In addition, we added further explanations and comments throughout the code sections to explain to readers what the code is being used for.

Comment 31: Also, the models using 'rethinking' would take a long time to run, so it would be nice to separate out the script for running the models from the script for processing the models (and save the output from the running models script, so people could just load the models if they didn't have time to run them).

Reply 31: We decided not to include the posterior samples from the Bayesian models because we think that these are not data but rather a part of the reproducible workflow. There are also no additional computations that occur with the posterior sample, we only report the compatibility estimates. While these models do take more time to run, we do not think that they are unreasonable: even the more complicated models can run on a desktop computer within a couple of hours. We have added a note about the run time to the script.

Comment 32: Fourth, there are just a lot of files in the repository, and there aren't clear instructions on how I, as an outsider, are supposed to use all of them, or what order they should be run in. For these reasons I did not check the computational reproducibility of the results in the preprint.

Reply 32: As mentioned above, the files on Github are mainly for convenience and we are referring to the archived versions for the data and code (see 'Data and Code Availability' section in our manuscript, DOI for previous version of the code deposited at the repository Edmond: https://dx.doi.org/10.17617/3.80). We now refer to these repositories wherever we previously only had a link to Github. For the Github repository, we have changed the folder structure in the repository and added an explanation to the readme on the landing page:

Github repository: "All code to repeat the analyses is provided in the Rmarkdown file of the post study manuscript (https://github.com/dieterlukas/FemaleDominanceReproduction_MetaAnalysis/blob/trunk/Manuscriptfiles/PostStudy_MetaAnalysis_RankSuccess.Rmd). The code in the file will automatically load all necessary input data from this repository. The folder 'Manuscriptfiles' contains the Rmarkdown, html, and pdf copies of the manuscript and the preregistration. The folder 'Inputfiles' contains copies of the data files".

DESCRIPTION OF LITERATURE SEARCH AND SCREENING

Comment 33: The study is presented as a "systematic assessment" (line 55), but the methods for finding and selecting studies are not described in sufficient detail to be considered systematic.

Reply 33: We had in the beginning constructed our literature search as a forward citation search, starting from published overviews on the topic. This appeared more feasible to achieve during the initial project phase, which started as an undergraduate project. The combination of this initial data collection, which was chosen due to the time constraint of having to finish a report for credit, with additional searches afterwards, means that we ended up not exactly following any of the established protocols. We now provide more details about our literature search (for more information see our replies to the following comments). With this in mind, we had meant to use the phrase 'systematic assessment' in the introduction not to refer to our data collection approach, but to the inclusion of multiple predictor variables and of their potential interactions. To avoid confusion we have now changed this:

Introduction: "However, there has been no study simultaneously examining the effect of life-history, social and ecological factors in modulating the benefits of social dominance. Similarly, there has been no quantitative assessment of the potential factors that may mitigate the relationship between rank and reproductive success to explain those cases where high rank is not associated with higher reproductive success."

Suggestions for improving the reporting:

Comment 34: (1) Provide dates for when searches were performed (i.e. to the day and month and year; July 2019-January 2020 is insufficient)

Reply 34: For all searches, going through the records and extracting the data took several days, so we had listed the general period. We have now specified this more clearly for each search to the day:

Methods: "(using Pubmed, 22 May 2019 - 13 June 2019). Next, we searched Google Scholar and Google Search with the following terms: "dominance AND female AND mammal AND reproductive success OR reproduction" (04 July 2019 - 31 July 2019; 143 additional effect sizes), "rank AND female AND mammal AND reproductive success OR reproduction" (14 September 2019 - 13 November 2019; 90 additional effect sizes), and "sex ratio AND dominance AND female AND mammal" (11 February 2020 - 06 March 2020; 75 additional effect sizes)."

Comment 35: (2) Provide details of how/where citations for the review papers were retrieved from

Reply 35: We added an explanation that we found the citations in the review papers from pubmed:

Methods: "We started with the references in previous major reviews and meta-analyses on the association between dominance and reproduction in female mammals (see below for inclusion criteria): @fedigan1983dominance (8 effect sizes on female primates entered), @ellis1995dominance (16 effect sizes entered / 5 not entered on female non-primates, 38 effect sizes entered / 22 not entered on female primates), @brown2002reconsidering (28 effect sizes entered / 7 not entered on female primates), @stockley2011female (12 effect sizes entered / 2 not entered on female non-primates, 11 effect sizes entered / 1 not entered on female primates), @majolo2012fitness (26 effect sizes entered / 2 not entered on female primates), @pusey2012magnitude (45 effect sizes entered / 2 not entered on female primates), and @clutton2013social (8 effect sizes entered / 1 not entered on female primates, 6 effect sizes entered / 1 not entered on female non-primates) (some effect sizes appear in multiple of these studies, leading to a total of 136 effect sizes) (using Pubmed, 22 May 2019 - 13 June 2019)."

Comment 36: (3) Provide the search strings for Google Schilar and Pubmed in their original form, with Boolean operators, saying which fields were searched

Reply 36: We have clarified how we combined the search terms we had listed.

Methods: "Next, we searched Google Scholar and Google Search with the following terms: "dominance AND female AND mammal AND reproductive success OR reproduction" (04 July 2019 - 31 July 2019; 143 additional effect sizes), "rank AND female AND mammal AND reproductive success OR reproduction" (14 September 2019 - 13 November 2019; 90 additional effect sizes), and "sex ratio AND dominance AND female AND mammal" (11 February 2020 - 06 March 2020; 75 additional effect sizes)."

Comment 37: (4) Line 308 states that only the first 1000 results were checked for all searches. What order were the 1000 results taken in? I'm guessing date for pubmed, 'relevance' for google scholar?

Reply 37: We used pubmed only to retrieve the references listed in the reviews/meta-analyses. We clarified this. For Google Scholar we indeed had the results sorted by 'relevance'.

Methods: "We limited our checks to the first 1000 results for all searches as automatically sorted by the respective search engine (sorted by 'relevance' on Google Scholar)."

Comment 38: (5) It's not reported how the studies were screened for inclusion. Two people did the literature search, did they also screen the studies? Was the screening divided between people, or was there parallel screening to check the agreement rate? Was there any pilot screening to measure the agreement rate of the inclusion/exclusion criteria?

Reply 38: Yes, we had pilot screened the articles in the meta-analysis by Majolo et al. We did meet and discuss throughout the initial data collection, but we had no further parallel screening. We have added this information.

Methods: "We initially coded eight papers independently, for which we both extracted the same values and classified the approaches in the same way. S and DL also independently went through the studies included in @majolo2012fitness and agreed on which to include and which not. After this, S and DL independently identified and coded articles, with occasional cross-checks and discussions of any border line cases."

Comment 39: (6) To clarify the processing of searching and screening studies, it would be useful to present the results in a PRISMA-style flow diagram (http://www.prisma-statement.org)

Reply 39: It is difficult for us to retroactively assign the filtering decisions to each of the steps in the PRISMA statement, in particular given the mix of approaches. We have tried to add the relevant information to the method section (see our replies 34-38).

ANALYSIS METHODS

Suggestions for improving the description of analysis methods:

Comment 40: (1) Reporting R version numbers and the names of functions used to run models (e.g. rma.mv)

Reply 40: For the R version, we have added this information, based on the main computer that performed the final analyses reported (S and DL had both run the analyses).

Methods: "We performed all analyses in the statistical software R (version 4.0.3; @Rlanguage)."

For the functions, we had assumed that interested readers would look at the Rmarkdown file, but have now also included this information in the main text.

" We first estimated all models using function "rma.mv" in the package metafor (@viechtbauer2010conducting)."

Comment 41: (2) As noted by the authors in their reply to the preregistration review, using 'rethinking' to run meta-analyses is very rare in ecology. Therefore, I would like more details on exactly what model was being implemented with this package, using mathematical notation, to compare-and-contrast the differences (if any, other than the estimation method) between the metafor approach and the model eventually fit within stan. I am not very familiar with the rethinking package (I confess to having started but never finishing the textbook!), so the code alone was not enough for me to tell what was going on "behind the scenes".

Reply 41: We have added more explanation to the implementation of the model for STAN:

Methods > Analysis: "We detail model construction in the following: we first assess whether species and population identity create dependencies amongst the measured effect sizes. If so, we include these factors through covariance matrices reflecting the dependence across measurements. The models take the following form: we assume that the transformed effect size 'observed Fisher Zr' we extracted from the articles (individual effect sizes indexed by [i]) is a reflection of the 'true Fisher Zr' effects that were measured with some error, the extent of which is reflected by the observed 'Variance' of each effect size; the 'true Fisher Zr' effect sizes come from an overall distribution, the mean of which depends on an intercept and the influence of the respective predictor variables; and where similarity in the variance of the 'true Fisher Zr' effect sizes from different species is reflected by 'sigmasquared' which is assumed to follow a Gaussian process with a multinormal prior that transforms the extent of the shared phylogenetic history among species pairs i and j with the parameters etasquared (covariance among closely related species) and rhosquared (decline in covariance as phylogenetic distance increases):

'observed Fisher Zr[i]' ~ normal('true Fisher Zr[i]' , Variance[i])

vector[Number of species]:'true Fisher Zr[i]' ~ normal(mu , sigmasquared)

matrix[Number of species:Number of species]: sigmasquared ~ MVNormal((0...0),K)

*K[ij]=etasquared times exp(-rhosquared * phylogenetic distance[ij])*

mu ~ alpha + beta*explanatory variable[i]

alpha ~normal(0,1),

beta*explanatory variable[i]~normal(0,0.5),

etasquared~Exponential(1),

rhosquared~Exponential(1)"

Comment 42: (3) It would be good to provide the equations to the effect size and its sampling variance. As noted above this is also something missing from the code, as pre-processed data were not provided.

Reply 42: As we described in the methods, we estimated the effect sizes using the online calculator at the Campbell Foundation. We often entered the observations from the respective articles directly in there. Accordingly, this is not part of our calculations here and therefore not in our code. We decided to refer to the references we used that contain the formulas.

Onto the presentation and readability:

REPORTING OF RESULTS

Comment 43: The results report the number of effect sizes, but given the high level of non-independence the number of studies should be reported at all instances where sample sizes are included too.

Reply 43: Our study is slightly different from other meta-analyses in that when studies report multiple effect sizes, they are usually reporting outcomes from different approaches (e.g. measuring interbirth interval and assessing offspring survival). Non-independence among effect sizes from the same study in our sample are therefore less likely to reflect methodological issues that were particular to a given study, than to occur because a study usually only focused on a single study system. In the models that include the random factors of study, species, population, and phylogenetic relatedness, the phylogenetic component (set to assume that values from the same species are similar) explains more than two times as much of the covariation among effect sizes as the study component. Other factors (e.g. whether a species is a cooperative breeder or not) explain even more variation. While we account for these interdepencies within studies, we are not sure that providing the information on the number of studies as sample size information for analyses is relevant information.

Comment 44: Also there is no mention of missing data in the preprint and how this was dealt with, and whether any of the planned analyses couldn't / shouldn't be run because of insufficient data.

Reply 44: We mentioned the one case where we could not run analyses because data were not available in the changes from the preregistration. We have now clarified this and added a detail about the procedure for cases where data on predictor variables was not available for all observations.

Methods > Preregistration: "we did not collect data on variance in relatedness because it was not possible to extract this information from most studies reporting relatedness levels"

Methods > Analysis: "For the predictor variables where we did not find data to match to all the observed effect sizes, we excluded the cases with the missing data from the analyses. We report the sample size for each analysis."

Comment 45: On the subject of non-independence, I don't think it's justified to run the "base model" with no random effect for study ID. There is far too much pseudo-replication (making the estimates overly precise, hence no prediction or confidence interval visible on Figure 1). Also, the funnel plot is not a meaningful tool for visualising publication bias with such high levels of non-independence.

Reply 45: Our study is slightly different from other meta-analyses in that when studies report multiple effect sizes, they are usually reporting outcomes from different approaches (e.g.

measuring interbirth interval and assessing offspring survival). Accordingly, non-independence here is less about methodological issues that were particular to a given study. In the models that include the random factors of study, species, population, and phylogenetic relatedness, the phylogenetic component (set to assume that values from the same species are similar) explains more than two times as much of the covariation among effect sizes as the study component. Accordingly, we think that in order to detect biases we need to initially focus on all the raw data. One of the biases can indeed be that even within studies researchers decided which aspects to focus on and/or which results they presented in the manuscripts.

We have now changed the model to derive the prediction and confidence intervals for Figure 1. We added a random factor "effect size id" to the model, to reflect that each effect size is a separate sample, which more accurately reflects the sampling variance.

Figure 1: "Overall, most studies report a positive association between dominance rank and reproductive success (darker circle in the center indicates the mean, thick black edge right next to circle indicates precision interval, thin black lines extending from darker circle the confidence interval of the estimate)."

READABILITY

Comment 46: I recommend adding a couple of sentences into the abstract: one summarising the main conclusions / take-home messages, and one summarising the limitations of this study.

Reply 46: We have added sentences to the end of the abstract to summarize our study (see also Reply 15):

Abstract: "Our findings indicate that obtaining a high ranking position in a social group consistently provides female mammals with fitness benefits, even though future studies might show lower effects given various biases in the literature we were able to access, including, but not restricted to, a publication bias. They further draw a complex landscape of the level of social inequality across mammalian societies, reflected by variation in the benefits of social dominance, which appears to be shaped by reproductive and social competition more than by ecological competition."

Comment 47: As a reader I might be more impatient than average, but I did find the preprint overly long and sometimes hard to follow. I ask the authors to consider whether a more traditional presentation of the introduction could help increase the readability. Currently we go from "Background", to "Objectives" (which included the four main predictions", and then to "Predictions" (where the predictions were broken down into sub-predictions). I suspect this is a carry-over effect from the way the pre-registration was written. In the pre-registration it is good to have all the predictions clearly laid out, but now that you have the results, I think it

would make more sense to present the predictions at the same time as presenting the results (and all predictions could be summarised in a table in the methods or something like that)? In the current form, it is a lot of front-loaded information, and many readers (including myself) won't remember the details of the various predictions by the time they get to the methods, let alone results, so it is tiring them out unnecessarily.

Reply 47: Thank you for the suggestion. The structure was indeed a hold-over from the preregistration. We have now changed the overall structure by removing the prediction section from the introduction and instead linking the predictions directly to the results. We agree that this helps to reduce repetition and can make the text more accessible.

REPORTING OF PRE-REGISTRATION

Comment 48: The title page states "The background, objectives, predictions, and methods are unchanged from the preregistration that has been pre-study peer reviewed", however there are deviations from the registered methods (specified in the section starting at Line 511, 'Changes for preregistration').

Reply 48: We had meant to express that the text in those sections was unchanged, but can see how this was confusing. In addition, we have now revised these sections based on the feedback, so we have changed the statement on the title page.

Title: "The background, objectives, predictions, and methods have been peer reviewed prior to analyses and received an In Principle Recommendation on 07 July 2020:"

Comment 49: The title page also gives the impression that pre-registration occurred prior to the study starting, but it was pre-registration prior to analysis (studies were collected, and a lot of data were collected, before preregistration)

Therefore would you be able to say something like:

"The background, objectives, and predictions were pre-registered prior to final data collection, and prior to any data exploration and analysis. Deviations from pre-registered methods are explained within the manuscript"

Reply 49: We have added the statement that:

Title: "Deviations from pre-registered methods are explained within the manuscript."

I think justifications are missing for these two deviations from the registration:

Comment 50: 1.   "We changed how we calculated sexual dimorphism in body weight."

Reply 50: We have added an explanation for this change (see also reply 2):

Changes from preregistration > additional variables: "We changed how we calculated sexual dimorphism in body weight. We had previously taken the ratio of male weight divided by female weight. A collaborator on a different project, in which we also use sexual dimorphism in body weight as a variable, alerted us to the article by @smith1999statistics which shows that this simple ratio is biased because its distribution across species is non-linear resulting in asymmetries when females are the larger sex (as example, assume a species where individuals of one sex are 10kg and individuals of the other sex are 8kg; if males are the larger sex the simple ratio would indicate that the larger sex is 25% larger [10/8=1.25]; however, if females were the larger sex it would indicate that the larger sex is only 20% larger [8/10=0.80]). We therefore switched to the calculation of sex ratio as recommended in this article by calculating sexual dimorphism as the average weight of males divided by average weight of females if males are heavier than females and as two minus the average weight of females divided by the average weight of males otherwise."

Comment 51: 2.   "We did not perform the multivariate analyses we had listed in the preregistration where the univariate analyses indicated no influence/interaction (group size + intersexual conflict; diet + population density; harshness + population density)."

Reply 50: Re-reading the plans for the multi-variate analyses we realized that we assumed that some of the interactions might indicate associations even if the analyses of the individual variables did not. We have therefore now added these planned multivariate analyses to the results:

Predictions/Results 5.4 - 5.9

OTHER COMMENTS

Comment 52: Figure 10 is hard to read with the overlapping colours… better to do density lines with no fill than histograms?

Reply 52: Thank you for the suggestion! We replace the figure with one using density lines, and agree that this makes it easier to see the distributions.

Figure 10: replaced

Comment 53: Lines 55-58: Found this sentence hard to understand. Break up into two?

Reply 53: We have broken up the sentence to better explain the idea (see also reply 16):

Background: "However, it has remained unclear whether and when there is selection on all females to compete for a high rank. Instead of direct selection on females to compete over high dominance rank because it provides substantial fitness benefits, selection might be on females to find a place in the hierarchy that maximizes their fitness based on their intrinsic qualities and access to social opportunities."

Comment 54: Lines 80-84:"we will perform" could now be changed to "we performed". Also, the second sentence needs to be broken up in some way, otherwise it reads as if your objective is your prediction (e.g. "…reproductive success. We predict…")

Reply 54: We changed the tenses throughout to past tense. We also reformulated this section to distinguish these objectives from the prediction:

Objectives: "In this study, we present a quantitative assessment of the strength of the relationship between female dominance rank and reproductive success in social mammals and explore factors that might mediate this relationship. Our objective is to identify the ranges of variation in the relationship between rank and reproductive success and to investigate how this relationship is influenced by differences in life-history, ecology, and sociality. We addressed our objective through the following questions, by testing the corresponding four core predictions, which each break into a number of secondary predictions (see results):"

Comment 55: Line 560: Figure 8a is mentioned here, but we don't see Figure 8 until much later in the manuscript.

Reply 55: To avoid this confusion we removed the reference to the funnel plot from this section and only refer to it later.

Comment 56: Lines 583-589: this paragraph is not easy to understand.

Reply 56: We have rephrased this paragraph to better explain the rationale and our interpretation of the result:

Results > Sample bias: "Studies with smaller sample sizes have a higher risk to report inflated effect sizes due to a higher likelihood of Type I and Type II errors. In our dataset, the average effect sizes at smaller sample sizes are more extreme than those at larger sample sizes (effect sizes range from -0.89 to +1.33 for studies with a sample size of 20 or smaller, while for studies with sample sizes larger than 20 they range from -0.37 to +1.24). However, it is not just that the spread of values is larger for studies with smaller sample sizes, but the positive bias in effect sizes we observe decreases with the sample size of studies (metafor estimate 95% confidence interval lower -0.03 to upper -0.02, rethinking estimate 89% compatibility estimate of posterior sample lower -0.09 to upper -0.04) (Figure 2). This

supports a 'publication bias', where studies with small sample sizes that did not show a positive effect are missing from the literature. However, the estimate of the intercept and slope of this model linking effect size to sample size shows that across the range of sample sizes the estimate of the overall effect size does not go below zero (see line in Figure 2). This indicates that females with higher rank have higher reproductive success across the range of sample sizes."

Comment 57: Also, presenting the confidence & credible intervals with a hyphen and a negative sign is a bit confusing (better with a comma than a hyphen).

Reply 57: We have changed the presentation of the intervals throughout the results.

"metafor estimate 95% confidence interval lower -0.03 to upper -0.02, rethinking estimate 89% compatibility estimate of posterior sample lower -0.09 to upper -0.04"

Comment 58: Line 1039: typo at "but not directly indicates"

Reply 58: We have rephrased the sentence:

Discussion: "First, it is important to highlight that this effect size reflects how well rank predicts reproductive success, but the effect size does not directly indicate how different the reproductive success of high-ranking females is from that of low-ranking females."