Dear Dr.'s Doligez, Anonymous reviewer, Slagsvold and Elliot:

Thank you so much for your thorough feedback on our manuscript! We really appreciate the time you have taken to give feedback. You have made this research better. We responded to your comments below and revised the manuscript accordingly (PDF version at XXXXX or HTML at http://corinalogan.com/Preregistrations/gcondition.html).

As this study was a pre-study peer-reviewed preregistration that received in-principal acceptance, we are not permitted to alter manuscript components that have already gone through this Stage 1 review. In the PCI Ecology guidelines for Reviewers at Stage 2 (post-study) it states: ****The reviewers from Stage 1 then return to assess the completed Stage 2 manuscript, focusing on compliance with protocol and whether the conclusions are justified by the evidence. Crucially, reviewers do not relitigate the theory, hypotheses or methods, thus preventing knowledge of the results from influencing recommendations. PCI editors (termed "recommenders") are similarly forbidden from rejecting a manuscript on the basis of any new concerns about the methodology or rationale, or on the basis of the main outcomes.****
(https://rr.peercommunityin.org/help/guide_for_reviewers#h_3015488595591613635204737). As you can see, these guidelines are important for decreasing the tendency to change hypotheses and the framing of the experiment after results are known. We recognize that this is a bit strange for you as none of you were reviewers in Stage 1 and we apologize for any confusion about this that may have led to additional effort. For detailed instructions on how to review a stage 2 registered report, please see the reviewer guidelines for the new Peer Community in Registered reports: https://rr.peercommunityin.org/help/guide_for_reviewers, PCI RR is replacing the system of peer reviewing preregistrations at PCI Ecology, which was the pilot program.

Note that the version-tracked version of this manuscript is in rmarkdown at GitHub: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/gcondition.Rmd. In case you want to see the history of track changes for this document, click the 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).

All our best,

Kelsey (on behalf of all co-authors)

Round #1

by ***Blandine Doligez****, 2021-04-15 21:30*

Manuscript: **10.32942/osf.io/857gt**

COMMENT 1: First of all, my deepest apologies for taking so long to answer, due to intense field work…

Three reviewers have now read your study and provided very detailed and constructive comments. They are different from those that reviewed your pre-registration (and I also did not handle your pre-registration as a recommender), but I find that they give highly relevant comments and hope that you will find them useful.

Basically, all were strongly interested in the study and acknowledge that the question is of high interest: making sure that indices widely used so far are valid measures of individuals' condition is a question of high importance yet most often overlooked. By aiming at testing whether this is the case, the study should provide highly useful general information to many other studies, and thus has a great potential to improve our understanding of previous results based on such indices.

Yet all three reviewers also raise a number of concerns, and I concur with them that these need to be addressed or at least discussed more thoroughly. As you will see, reviewers had many common comments, which means that many readers will also ask themselves the same questions, which need to be clarified. Among the major points raised:

>**RESPONSE 1:** We appreciate the time and effort that you and the 3 reviewers committed to improving this manuscript. We address the comments below.

COMMENT 2: whether the species is ideal to test the proposed hypotheses may not be that obvious, especially when it comes to estimating reproductive success and the reason for nest failure (as noted by authors themselves) but also given the limitation of sample size or physiology in arid zones (see reviewer's detailed comments). More nuance may be needed with a balanced presentation of the advantages but also drawbacks of conducting this study on this specific species. This comes along with more information needed about the species biology (in particular to assess the reproductive success measures).

>**RESPONSE 2:** We appreciate the concern about the appropriateness of addressing this question with great-tailed grackles. We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change that now.

However, we added more information to the discussion about grackle biology compared to other bird species, and the benefits and drawbacks of focusing on this species for this question:

"Great-tailed grackles are an interesting system to study energetic condition and reproductive success because they rapidly expanded their range into Arizona, where the climate and habitat are distinct from that in Central America where the species originally evolved [@wehtje2003range]. The increase in temperature variation and decrease in available water are both environmental stressors that have previously been found to negatively affect energetic condition [@pendlebury2004variation]. Reproductive success is vital to species persistence and abundance in novel environments [@masponsbehaviour2019]. Therefore, an understanding of energetic condition and its association with reproductive success in grackles outside of their original range could broadly inform conservation research in invasive and non-native species. While the reproductive success of certain avian species may be easier to monitor at a more fine scale (i.e. cavity nesters), the predominant measure of reproductive success currently used by avian ecologists is the ability of adults to fledge offspring [since foundational work by @mayfield1961nesting] because it is financially and logistically accessible to more researchers. Therefore, our measure of reproductive success in grackles should be informative, and research that spans taxa with diverse reproductive strategies is important for understanding general trends in energetic condition and appropriate proxies."

COMMENT 3: the need for more than just two coarse indices to be used to reach a full validation (at least linking them with more direct measures of energetic state would be really needed), as well as the need to account for temporal variation in these indices along the day and the season, which was not done here.

>**RESPONSE 3:** We found no evidence for temporal variation at the scale of season. It was logistically impossible to assess temporal variation at a finer scale, however we received pre-study reviews and in-principal acceptance with this study plan, so we cannot change that now.

COMMENT 4: discussing what energetic condition represents really, and whether this can be considered as a 'single' trait or whether it should be a multi-facetted trait for which different proxies will measure different facets (and thus be potentially uncorrelated).

>**RESPONSE 4:** The aim of our article was not to distinguish whether the inherent energetic condition phenotype should be considered as one or multiple traits. Current research in the field often simplifies energetic condition to just one trait and attempts to measure it with a particular method (Green 2001; Labocha et al. 2014; Barnett et al. 2015). This is particularly true for the morphological proxy measures that we focused on here (in contrast to physiological proxies like measures of immunity). Therefore, we aimed to challenge the assumption that multiple proxy measures equally reflect energetic condition when they have not explicitly been compared.

We added the following to the discussion to address this point:

"Energetic condition is not directly observable, but variation can affect life history characteristics [@labocha2014which; @barnett2015mass]. Consequently, a large corpus of research attempts to measure energetic condition using various proxy measures [@labocha2014which] and largely assumes that the chosen proxy accurately reflects energetic condition as a singular trait. Although it is often implicitly assumed that all proxy measures for energetic condition reflect the same inherent trait, it is rare for one study to compare multiple proxies. However, if all proxy measures are affected similarly by a singular energetic condition phenotype, then multiple proxy measures will produce correlated results. The aim of the current study was therefore to test the idea that multiple commonly used morphological proxies equally measure energetic condition and that the measures can explain variation in life history characteristics."

Green AJ. 2001. Mass/length residuals: measures of body condition or generators of spurious results?. *Ecology 82(5)*:1473-83.

Labocha MK, Schutz H, Hayes JP. 2014. Which body condition index is best?. *Oikos 123(1)*:111-9.

Barnett CA, Suzuki TN, Sakaluk SK, Thompson CF. 2015. Mass‐based condition measures and their relationship with fitness: in what condition is condition?. *Journal of Zoology 296(1)*:1-5.

COMMENT 5: explaining better the mechanisms underlying the hypotheses and expected relations between variables (in particular with reproductive success)

>**RESPONSE 5:** We added the following to the discussion to address this point:

"Energetic condition can have a large impact on reproductive success in birds [@montreuil2017relationships, @drent1980prudent] and flying mammals [@welbergen2011fit]. For example, female chickadees with higher winter fat scores are more likely to lay eggs earlier in the subsequent breeding season, as well as go on to feed those offspring more frequently [@montreuil2017relationships]."

COMMENT 6: the power of the analyses. This is particularly true regarding the use of the fat score, **which had to be turned into a binary variable due to very limited variability**, and for its relation with reproductive success – from my own experience, it appears that reaching enough power when relating two binary variables as here can require really high sample sizes! Globally, power analyses suggest that the sample sizes obtained here will probably not be enough to reach sufficient power. While it seems that more data cannot be obtained (due to the end of the field work in the study area), this point may need more discussion: what is to be concluded from the results here in the end?

>**RESPONSE 6:** We received pre-study reviews and in-principal acceptance with this sample size, therefore sample size cannot be a reason for critiquing this stage 2 manuscript. However, since converting fat score to a binary variable was a deviation from the preregistration, we consulted with statisticians. They did not believe there was any reason that relating two binary variables would require significantly greater sample sizes compared to relating a binary response to a continuous predictor. Could you please share a citation for your comment?

COMMENT 7: a focus in the introduction on species that share energetic characteristics with grackles (i.e. endotherms and flying organisms).

>**RESPONSE 7:** We received pre-study reviews and in-principal acceptance with the introduction as is, so we cannot significantly change that now. However, we have added the citations suggested by Reviewer 3, and we have expanded the discussion, which now includes our results in context with results on other flying organisms:

“While both proxies are well supported in previous research as measures of energetic condition, our results indicate that they may not be measuring the same trait. This has also been found in studies on bats [@mcguire2018common], which are species that similarly experience distinct demands on body structure to facilitate flight.“

AND:

“Energetic condition can have a large impact on reproductive success in birds [@montreuil2017relationships; @drent1980prudent] and in flying mammals [@welbergen2011fit]. For example, female chickadees with higher winter fat scores are more likely to lay eggs earlier in the subsequent breeding season, as well as go on to feed those offspring more frequently [@montreuil2017relationships].”

COMMENT 8: I add below a few more comments:

- abstracts usually have no references; although there is no theoretical reason not to include some there, maybe it would be better to conform to a standard paper format?

>**RESPONSE 8:** Thank you for this feedback, we removed the citations in the abstract.

COMMENT 9: beginning of the abstract: why focusing on morphological traits? Later on you include hematological indices, which are physiological traits. Maybe start more generally here (e.g. on top of physiological variation, behavioural variation also has the potential to influence life history characteristics, and may also potentially be used as condition indices)

>**RESPONSE 9:** We understand the confusion here and we revised the manuscript to distinguish when we are referring to morphological or physiological proxies throughout. We focused on

morphological proxies because these are the predominant methods used to estimate energetic condition in birds in the field so far. We removed the mentions of physiological proxies from the abstract and only mention these in the discussion of future research directions.

COMMENT 10: in the introduction, the comment about reproductive success being difficult to monitor is strongly dependent on the species considered; while it may apply here (but not totally coherent with what is then described in methods) and to open-nesters in general, this is far less true for cavity-nesters or colonial birds, with many studies collecting routinely such data. This part may thus need to be nuanced, and not presented as applying in general. (This also relates to the point whether the study species is really ideal to investigate this question, as pointed out by reviewers). See also the end of the discussion, for which the question of not detecting nest failures is rather specific and does probably not apply to most species. Furthermore, **the reference justifying the predominant use of fledging success as a measure for reproductive success as a whole is quite old** and more discussion based on recent reviews is needed here.

>**RESPONSE 10:** We received pre-study reviews and in-principal acceptance with the introduction as is, so we cannot significantly change that now.

Additionally, please see Response 21, below.

Lastly, we used the Mayfield 1961 reference to justify using fledging as a measure of reproductive success because this is a foundational paper in the field. It introduced a now popular and well-supported data analysis technique for nest success that relies on the number of days until a nest fledged offspring or failed. We are unaware of more recent reviews that support an alternative measure of reproductive success in birds.

COMMENT 11: the end of the introduction should provide clearer information about what has been done in the study; **maybe include a shorter version of the subsequent hypotheses, predictions and alternatives, currently very long and detailed**?

>**RESPONSE 11:** We received pre-study reviews and in-principal acceptance with the introduction as is, so we cannot significantly change that now. However, we have revised the subsequent section to consist of abbreviated hypotheses, and added a summary methods section above the results, which is a more typical style in journal articles.

COMMENT 12: in the ‘deviation from the planned methods’ section, repeatability analyses are evoked, as well as data distributions, but we have no methodological information beforehand on why these analyses were done / what data exactly is used. Maybe it would be more logical to present the methods first before pointing out the changes that had to be made compared to the pre-registration (otherwise these details are not known before). In other words, the current order of sections does not appear the most logical.

>**RESPONSE 12:** See Response 11.

COMMENT 13: some sections repeat themselves (e.g. the calculation of SMI, which should be done in Methods rather than in Introduction and Results; categorization of SMI, etc.) – see also the remark below on the general presentation

>**RESPONSE 13:** Thank you for this feedback. We have removed the calculation of SMI and the description of the SMI categories from the Results. Because we received pre-study reviews and in-principal acceptance with the introduction as is, we cannot significantly change that now.

COMMENT 14: wing length could be argued not to represent skeletal size better than tarsus length… Another way to choose the variable measuring size could be the variable with the highest repeatability instead. From my own experience, I would say that a repeatability of 0.5 for female wing length does not seem very high.

>**RESPONSE 14:** We can see how this may not be intuitive. However, wing length and tarsus length are used interchangeably in the literature to represent avian skeletal body size. Wing length and wing surface area in birds is an important component of structural body size because increased mass requires larger wings to allow for flight, so the two measures should be correlated.

We stated in our preregistration that we would choose the measure most correlated with body mass, so we cannot change to tarsus length now. Furthermore, the repeatability value of tarsus length in females was smaller than that for wing length (0.4) and in males the two repeatability values were very similar (0.75 and 0.76).

COMMENT 15: provide statistics for the test of the non-linear relationship between SMI class and reproductive success (currently only presented as a figure, Fig. 3?). Also, as noted by reviewers, provide all statistical information, and not just p values.

>**RESPONSE 15:** We apologize for the confusion associated with this component. Due to our small sample size, before adding an additional covariate representing a non-linear effect of SMI, we proposed to first visually assess whether there was evidence that this was occurring in our data. We updated our phrasing in Results > P2:

"To determine whether we should include any non-linear effects of SMI in our models [@milenkaya2015body; @gosler1995predation], we visually evaluated whether individuals in any of 5 categories, ranging from low to high SMI, were more likely to be reproductively successful (Fig. 2). We found no visual evidence for a non-linear relationship between reproductive success and SMI for males or females (Fig. 3). Consequently, we did not include non-linear terms in subsequent models."

Please see Response 55 below for how we included additional statistical information in the Results section.

COMMENT 16: methods: regarding independent variables for P2: why choosing to use fat score, i.e. a binary variable, rather than SMI, i.e. a continuous variable, if both are correlated? I would expect the contrary, to increase the chances of detecting correlations / effects (at least you should justify this choice).

>**RESPONSE 16:** We tested whether SMI and fat score were correlated and we found that they were **not correlated**. Therefore, we included both as independent variables in our model testing P2.

COMMENT 17: overall, the general presentation of the study was surprising to me (detailing hypotheses and predictions, and detailing methods as done). Maybe a more 'standard' presentation would make it easier to grasp an overview of the study and its implications…

>**RESPONSE 17:** See Response 11.

Review by anonymous reviewer, 2021-02-12 16:32

COMMENT 18: This study sets out to test how well morphological condition indices describe "condition" defined as "energetic condition" (or energetic state) and how well those indices relate to performance in terms of reproductive success.

For this two available and widely used morphological measures are used (visible subcutaneous fat, an index based on body mass and body size) and available measures of reproductive success (female nest success, male holding a territory with nests).

The approach to evaluate and validate the use of widely used morphological indices, and to evaluate multiple indices, is valuable and welcomed. While I was enthusiastic when I read the abstract I was later getting confused by how the study done not quite matched the aims (validation; at least not the way I would expect validation to be done) and the weakness of some of the measures used (coarse measures, see detailed comments below), small sample sizes, lacking information on distribution of some variables, as well as lacking information about species biology, that make it all difficult to draw any clear conclusions. Overall I am left with too many question marks. See detailed comments below.

>**RESPONSE 18:** Thank you for the thorough review of our manuscript. Below, we addressed the issues you point out. However, as noted above, because this is a review of a Stage 2 manuscript we are unable to change aspects of the manuscript that have already undergone review at Stage 1 and received in-principal acceptance (e.g., sample size, choice of variables).

COMMENT 19: Main issues I have:
The authors argue that the study species is an ideal system to study this – as reason the state that grackles have expanded and, being a water associated species, in the study area breed in an arid conditions which is assumed to bring physiological challenges. Hence expecting energetic condition really to make a difference for individual performance, as measured by e.g. reproductive success. While I can agree to this I doubt, however, that the species, or rather the study, is ideal:

>**RESPONSE 19:** Please see Response 2 and Response 21.

COMMENT 20: The proxy measures (indices for condition) are coarse. Surley you want to evaluate these course measures – but for an ideal study it would be good with some **more direct measures of energetic condition** so you can also test how well the coarse measures actual measure what we are set out to measure (this should be part of the validation, see ecological validity in Reale et al. Integrating animal temperament within ecology and evolution https://onlinelibrary.wiley.com/doi/full/10.1111/j.1469-185X.2007.00010.x).

>**RESPONSE 20:** The goal of our article was to test whether non-destructive indirect proxies commonly used in the field are measuring the same trait, and whether that trait is associated with reproductive success. There is evidence that lipid deposits are the major energy reserve for birds (Griminger 1986; Witter & Cuthill 1993), and it was not our goal to more directly measure lipids through destructive measures that few researchers want to undertake for their studies. Furthermore,

we received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that the results are known.

Please see Response 4 above for how we elaborated in our discussion on the nature of energetic condition as being one or multiple traits, and the relation of that topic to the aim of our article.

Griminger P. 1986. Lipid metabolism. *In* Avian Physiology (ed. PD Sturkie), pp. 345-358. Springer, New York.

Witter M and Cuthill I. 1993. The ecological costs of fat storage. Philosophical Transactions of the Royal Society: Biological Sciences 340(1291):73-92.

COMMENT 21: Furthermore, the measures of reproductive success are coarse, and it is also hard to know what they say because **we do not know the cause for nest failures**. If nestlings die due to rainy conditions or predation this may have little to do with the parents' energetic condition. Information about causes of nest failure seems quite crucial when using nest success as a measure for reproductive success that may be linked to energetic condition of the parents. It would also be better with being able to monitor number of nestlings as this would allow relating this to the parents' ability to provision them. You may not have information on causes of nest failure or number of fledglings due to methodological difficulties, but without this information your study system is not quite ideal for the questions studied.

>**RESPONSE 21:** We received pre-study reviews and in-principal acceptance of our reproductive success variable, so we cannot change it now that results are known.

Given that all individuals at our relatively small study site (<2.7 km^2) are experiencing the same environmental conditions, we can interpret the success of one nest and failure of another to be due to something other than environmental conditions. Energetic condition can also be related to variation in nest predation. For example, energetic condition has been shown to be related to nest predation via 1) variation in mobbing to deter nest predators (Abolins-Abols & Ketterson 2017) or 2) duration that the female can go without food during incubation to reduce the chance predators can detect and access the nest (Gill 1994; Fontaine & Martin 2006). Measuring the exact reasons for nest failure, the number of nestlings or fledglings, etc. is a massive logistical (monitoring many nests every day) as well as financial (cannot be accomplished without nest cameras; Williams & Wood 2002; Andes et al. 2019) undertaking for studies on any bird species and was outside of the scope of this study.

Gill FB. 1994. Ornithology. 2nd ed. WH Freeman and Company, New York.

Williams G & Woods P. 2002. Are traditional methods of determining nest predators and nest fates reliable? An experiment with wood thrushes (*Hylocichla mustelina*) using miniature video cameras. *The Auk 119(4):* 1126-1132.

Fontaine J & Martin T. 2006. Parent birds assess nest predation risk and adjust their reproductive strategies. *Ecology Letters (9):* 428-434.

Abolins-Abols M & Ketterson E. 2017. Condition explains individual variation in mobbing behavior. *Ethology (123):* 495-502.

Andes A, Shaffer T, Sherfy M, Hofer C, Dovichin C & Ellis-Felege S. 2019. Accuracy of nest fate classification and predator identification from evidence at nests of Least Terns and Piping Plovers. *Ibis, 161(2)*: 286-300.

COMMENT 22: This brings me to another question: I am confused about methods. In the Disc you state that you cannot determine causes of nest failures because you cannot access nests (nest are high and well concealed). But this statement conflicts with that you seem to have number of fledglings (http://corinalogan.com/Preregistrations/g_withinpop.html).

>**RESPONSE 22:** The withinpop.html preregistration was not a peer reviewed preregistration, but rather a place where we listed all of the variables we thought we would be able to collect data on and what we would predict about particular relationships between these variables. It turns out that we were not able to collect data on as many variables as we had initially planned because the nests were so high that we were not able to reach them to band nestlings, or count the number of eggs in the nest, or the number of nestlings, etc. We were occasionally able to see one or more fledglings in the outer branches of the nest tree, but more frequently they are observed being fed by the parents outside of the tree. Additionally, using fledglings as our indication of a female's reproductive success is appropriate because, by definition, fledglings move away from the concealed nests so they are more likely to be observable at some point. Whereas, nest concealment makes it impossible to determine the cause if a nest fails in the building, laying, incubating or nestling stage. However, we do not believe we should use the number of fledglings as our reproductive success variable because in nearly all cases we cannot be certain of the exact number of fledglings. Fledglings were not color banded, so it is difficult to tell them apart when they are mobile and the mothers may have fed additional fledglings outside of our view. Lastly, even in species that have nests that are visible, the nests can fail for reasons that are not visually detectable by researchers (i.e. hatching failure due to inbreeding or incubation failure; Gill 1994).

COMMENT 23: Some other difficulties I have with the manuscript: One of your stated aims is to investigate whether the two proxies measure the same trait (Abstract: "rarely there have been direct comparisons among proxies to validate that they measure the same trait", "validate whether they measure the same trait") – what is meant by the "same" trait? To me this does not make sense. (energetic) **condition is unlikely just one trait but rather a composite** – a number of traits acting together to shape energetic condition. So when talking about "condition" multiple measures unlikely measure the same trait – rather it amy make sense to use multiple measures/proxies to capture multiple aspect of "condition". So that for me the question is not whether visible fat and SMI measure the same trait, but rather how well they measure something that relates/contributes to energetic condition. Of course there might be different views on this – but i**mportant here is to clearly explain WHY you want to test whether the two proxies measure the same trait**. At present it's not clear.

>**RESPONSE 23:** Please see our Responses 4 and 20 above.

COMMENT 24: Fat score – relation to reproductive success: I am not sure that the assumed relationship fat score – condition – fitness is very relevant for reproductive success. While fat score shows strong links to migratory performance and survival during migration – the links to reproductive performance are not so clear or well established. Many bird species show very little or no visible fat during breeding (lean). We can assume that energetic condition plays some role for parental performance and hence may affect reprod success, but it is perhaps not that obvious that this energetic state can be described by visible subcutaneous fat (fat scores). So for me fat score seems a poor proxy.

>**RESPONSE 24:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known.

However, in response to your uncertainty about the relationship between subcutaneous fat and reproduction, please see Reviewer 3 comment 86 and associated citation, below. There is quite a lot of evidence that measures of body condition (including subcutaneous fat) vary in relation to the breeding season and reproductive success (Kitaysky et al. 1999; Milenkaya et al. 2015), indicating that there is some benefit to having more (or fewer) fat stores during specific stages of the annual cycle.

Kitaysky AS, Wingfield JC, and Piatt AF. 1999. Dynamics of food availability, body condition and physiological stress response in breeding Black-legged Kittiwakes. Functional Ecology 13(5): 577-584.

Milenkaya O, Catlin DH, Legge S, and Walters JR. 2015. Body condition indices predict reproductive success but not survival in a sedentary, tropical bird. PLoS One 10(8): e0136582.

COMMENT 25: The fat scoring has been developed for measuring visible fat when birds put on extra subcutaneous fat for e.g. migration, but fat scores may not work so well to measure every day fat/condition. What is the evidence in the literature?

>**RESPONSE 25:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known. It was not logistically feasible to measure fat score in our subjects every day. Furthermore, we were not interested in daily fluctuations in fat, but rather the average relationship of subcutaneous fat to reproductive success and SMI.

COMMENT 26: And similar for the SMI – it would be good to know more about the mechanistic background to see how body mass or a SMI can be closely related to variation in energy condition that would relate to parental performance and hence to reproductive success.

>**RESPONSE 26:** Please see Response 5 above for how we expanded on the relationship between energetic condition and reproductive success in the discussion.

COMMENT 27: While I agree that is is important to evaluate the use of the chosen proxies, for an ideal study I would expect more proxies to be tested, and a more complete validation including the links to energetic state (physiology).

>**RESPONSE 27:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known. As in most studies, the ideal design is often logistically and financially prohibitive, but incremental knowledge can still contribute to an understanding of the question at hand. Please see Response 2 for further details.

COMMENT 28: I am also lacking background (description of species biology) for why the reproductive success measures were chosen/used. Males holding territories with nests: what do you know about what traits/behaviours relate to a male's ability to attract a female and initiate nesting in his territory? Competitive ability is likely relevant and this may have to do with size (but perhaps not so much with visible fat), but a male's ability to hold a terr with nests may also depend on other behaviours/performance that do not necessarily have to be linked to body mass.

>**RESPONSE 28:** We apologize for the insufficient background information on the species reproductive ecology. Previous research in this species has shown that larger and heavier males were more likely to hold territories, have more social mates, and sire more offspring (Johnson et al. 2000).

We updated the discussion section as follows:

"...previous research in great-tailed grackles found that larger and heavier males were more likely to hold territories, have more social mates, and sire more offspring [@johnson2000male]. Our study additionally considered female morphology and reproductive success, subcutaneous fat, and controlled for the impact of structural body size on mass."

"We additionally used logistic exposure models to determine whether the energetic condition of females related to the probability of daily nest survival. **We only looked at females in this analysis because males were never observed contributing to nest building, incubation, or feeding nestlings in our population and so will not have a direct effect on daily nest survival.**"

COMMENT 29: Validation: For the indices to be valid proxies of condition we assume a relation between proxy – condition, a relation between condition – reproductive success, and also a relation between proxy – reproductive success. For a complete validation, you would want to evaluate all three relations, but here you only test proxy – reproductive success.

>**RESPONSE 29:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known.

Additionally, please see Response 20.

COMMENT 30: Ecological validation as suggested by Reale et al (see above) involves two main steps: 1) correlation to traits it is supposed to measure – i.e. corr to some measure of energetic condition and/or **things that are strongly related to energetic cond e.g. flight endurance**, 2) corr to fitness. Here you test 2, but part 1 (important part of validation: how well do morph traits reflect energetic condition?) is missing.

>**RESPONSE 30:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known.

Our understanding of the idea of measurement validation comes from Carter et al. 2013, a paper that follows up on Reale et al. 2007 and gives more explicit guidance on using proxy methods for measuring inherent traits. "Validity refers to the degree to which a test measures the targeted trait (Burns, 2008; see also Reale et al., 2007). A frequently recurring critique of personality psychology involves the validity of the tests that are used to measure the trait of interest (Duckworth & Kern, 2011). One of the first ways to remedy this problem is to use multiple measurements for multiple traits, and investigate correlations among the measurements (Campbell & Fiske, 1959)."

Carter AJ, Feeney WE, Marshall HH, Cowlishaw G, Heinsohn R. 2013. Animal personality: what are behavioural ecologists measuring? *Biological Reviews 88(2)*:465-75.

COMMENT 31: If the result of the study is no correlation between morphological trait and fitness: is this because of the trait being a poor proxy for energetic condition or because energ cond does not affect fitness much in the investigated context (e.g. fat score may be very strongly related to survival during parts of the year (migration) but be poorly related to reproductive success)?

>**RESPONSE 31:** This is a good point and we have now elaborated on this in the discussion as follows:

“Additionally, studying traits that could relate to variation in energy stores, such as dispersal [@ellers1998field], migratory endurance, [@deppe2015fat] or survival [@liao2011fat] would allow us to disentangle whether morphological proxies like fat score and SMI are poor proxy measures for energetic condition, or whether fat score and SMI do not affect reproductive success but may be associated with other life history characteristics.”

COMMENT 32: You state yourself in the Discussion that it is important to “ensure each proxy is measuring the intended trait”. If the intended trait is “energetic condition” then we want to know that the proxy can measure this and we can only get to know this by relating the proxy to one or more traits that are an expression of energetic condition (physiol + behavioural traits relating to performance, endurance). But that part is missing here.

>**RESPONSE 32:** We received pre-study reviews and in-principal acceptance with this study plan, so we cannot change it now that results are known.

Please see Response 30 for a description of our understanding of the idea of proxy validation. Our motivation for this study was to correlate proxy measures of energetic condition commonly used in this field, as there is a broad lack of support for what method does actually measure “true” energetic condition. It is outside of the purview of this article to relate energetic condition to traits other than reproductive success, like endurance, that theoretically should be directly impacted by energetic condition. We added migratory endurance to the future directions piece of the discussion copied above in Comment 31.

COMMENT 33: Related to this: here http://corinalogan.com/Preregistrations/g_withinpop.html I can see that you (at least sometimes) measure (apart from number of eggs, number of fledlings) also some immunological measures. Why are you not using those here?

>**RESPONSE 33:** We aimed to compare two of the most commonly used proxies for energetic condition to inform the measures that are most likely to be used in future research. Gathering immunological data (like blood hematocrit) can be logistically and financially costly, so fewer researchers are able to incorporate these data. Secondly, recent research shows that immunological measures like hematocrit are poor indicators of energetic condition (Fair et al. 2007).

You are correct that we do have immunological measures, but unfortunately now that our results are known we cannot add additional analyses that include these variables (also the blood has not yet been processed to quantify immune measures so these data are not available to us). Our study, as is, was pre-study peer reviewed and received in-principal acceptance, therefore we only need to conduct the work we proposed to do to complete this piece of research. In the future, others would be welcome to use our published data on more variables to conduct further analyses.

Fair J, Whitaker S, Pearson B. Sources of variation in haematocrit in birds. Ibis. 2007 Jul;149(3):535-52.

COMMENT 34: In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis). Yes, but no discussion about that the actual sample size is well below the sample size calculated as required for ca 80% or 90% power.

>**RESPONSE 34:** We understand the concern about sample size. However, our planned sample, which closely matches our actual sample size, and the power analyses were reviewed by multiple reviewers and the recommender and accepted by PCI Ecology in Stage 1. Furthermore, our sample size is in line with many other studies in this field because in many systems a sample of 89 or more individuals is logistically prohibitive.

COMMENT 35: Discussion: The conclusions are supported by the results, but **I would argue that results are weak (small sample, low power, poor/coarse measures**). While the overall conclusion (we need more validation and future research to analyse some relationships in more details) are supported, the expectations for a valistaion of the proc´xies are not fulfilled.

>**RESPONSE 35:** We are sorry that you feel like our manuscript did not meet your expectations. We hope that you will continue to engage with PCI Ecology and PCI Registered Reports to share your input with researchers during Stage 1 when these kinds of methodological changes can be made.

We are glad that you think our conclusions are supported by the results, which is the necessary component for acceptance of our manuscript at Stage 2

COMMENT 36: Condition/body condition/energetic condition: these terms are used interchangeably. Since you define “condition” in this study as energetic condition it would be good to be clear throughout the manuscript and always use the same wording “energetic condition”. In particular – avoid using “body condition” which is much broader (this term pops up in the Disc).

>**RESPONSE 36:** Thank you for this input! We have changed the wording throughout to always use “energetic condition”.

COMMENT 37: “one morphological trait ... is energetic condition”: It may be semantic but I find it difficult to view “energetic condition” as a morphological trait, or as one trait. To me energetic condition is a state that depends on a number of traits, and most of those are more physiological.

>**RESPONSE 37:** While physiological traits like immunity can be included under the umbrella of “body condition”, we focus here on the more commonly used morphological traits thought to indicate condition (fat, body size and mass). We understand your point though and throughout we have adjusted our wording to clarify when we are talking about energetic condition proxies that are morphological (fat score, SMI) or physiological (immune traits).

COMMENT 38: “condition proxy might relate ... in an unexpected way”: why unexpected? To me it’s not unexpected that “too high” reserves may have some negative consequences.

>**RESPONSE 38:** Thank you for this feedback, you are correct that there is evidence that too much fat can be costly (e.g., McNamara et al. 2005). We’ve updated the manuscript as follows (changes indicated in bold font):

“In some instances, the condition proxy might relate to life history characteristics **but in a non-linear way.**”

COMMENT 39: “Multiple proxies should correlate with each other if they measure the same trait (energetic condition).”: To me this does not make much sense because **I view energetic condition not as one trait**, and multiple proxies may not be correlated if they measure different complementary aspects (traits?) relating to an individual’s state described by “energetic condition”. Maybe it’s just me missing something – but the way it’s written I do not understand why I should expect that multiple proxies to be correlated. I might even want multiple proxies with little correlation that describe several complementary aspects that together give a better description of the state then several correlated proxies?

>**RESPONSE 39:** Please see Response 4 above.

COMMENT 40: “we compare two indices (fat score and the scaled mass index) of an individual’s energetic state to validate whether they correlate with each other, which would indicate that they both measure body condition": since I think of body condition (or energetic condition) as being determined by several traits that can be complementary (with potentially little correlation) this aim does not make sense. At least not without better explanation.

>**RESPONSE 40:** Please see Response 4 above.

COMMENT 41: Parts of the text in the Intro would be better placed in Methods, e.g. details on fat score, details in SMI calculation.

>**RESPONSE 41:** We received pre-study reviews and in-principal acceptance of the introduction in the current state, so we cannot change it now.

COMMENT 42: Instead of details about the scores I would rather want information about what is known how fat score relate to condition, or e.g. every-day performance of behaviours. For example, "is ... a method for estimating condition": in which situation? What kind of studies? Only during migration, or winter, or also during the breeding season?

>**RESPONSE 42:** We received pre-study reviews and in-principal acceptance of the introduction in the current state, so we cannot change it now.

However, please see Response 26 above for how we elaborated on this in the discussion.

COMMENT 43: You cite a study by Haas 1998 that did not find a relationship between fat score and nest success – but I also would like to hear about studies that do so. You state "previous research found that it does not always positively relate to life history variables" – that means that is sometimes does – but you don't give any examples. So I am left with no information about the actual evidence of fat scores ...any relationship to ...condition-affected life history or behavioyral traits (other than migration).

>**RESPONSE 43:** Please see Response 26 above for how we elaborated on this in the discussion.

COMMENT 44: Body mass: was this controlled for time of the day? What do you know about how body mass varies across time of day in grackles? Or how it varies with the amount of food recently eaten? How does it vary with time of year (across seasons)? In relation to moult? Could such variation cause noise in your measures making it more difficult to detect relationships with other variables?

>**RESPONSE 44:** We measure wild-caught grackles, so it is impossible to know what amount of food was recently eaten. SMI was highly correlated with body mass and our results show that it did not vary across season. It was logistically impossible to measure grackle weight multiple times per day or even more than once because, in most cases, we are not able to retrap grackles in the wild. We received pre-study reviews and in-principal acceptance of these methods, so we cannot change it now.

COMMENT 45: Lacking information about species: what motivates the use of whether a male defended a territory containing nests or not as a proxy for reproductive success in males? I guess you can only use such a measure if there is variation such that some males attract one or more females and some do not, but **I would like to get some information about this, and what the normal proportions of males with or without nests are**. And what kind of resources, what traits are important for males to succeed attracting females that build nests?

>**RESPONSE 45:** In an in-depth assessment of mating strategies of great-tailed grackles, Johnson et al. 2000 found that one-third of color-banded males held a territory containing nests. Territorial males were larger, heavier and had longer tails.

Please see Response 57 for the added information on the proportion of our sample that held territories.

We added this additional background information on typical male breeding ecology to the discussion as follows:

"Energetic condition is likely a factor in reproductive success in our system because **previous research in great-tailed grackles found that larger and heavier males were more likely to hold territories, have more social mates, and sire more offspring [@johnson2000male]. Our study additionally considered female morphology and reproductive success, subcutaneous fat, and controlled for the impact of structural body size on mass.**"

COMMENT 46: Prediction 2 alternative 2: no correlation can also be resulting from small sample sizes (low power), too course measure of nest success,

>**RESPONSE 46:** We received pre-study reviews and in-principal acceptance of the predictions in their current state, so we cannot change it now.

COMMENT 47: Fat scores: majority of score were zero – what was the exact proportion? This is important to know, because if the proportion is very large there is hardly any variation that can be used to analyse relationship to either SMI or reproductive success. Please show data how many of the 21 male fat score values were 0 or 1, how many of the 47 female values?

>**RESPONSE 47:** Thank you for this feedback. We added the proportion of zeros to the Deviations from the planned methods section:

"...Specifically, of the 21 males, 15 had fat scores at 0, 5 scored 1, and a single male had a fat score of 2. Out of 47 females, 26 scored 0, 18 scored 1, 2 scored 2, and a single female scored 3."

COMMENT 48: "We were able to calculate SMI for 24 males and 62 females, and fat score values were available for 21 males and 47 females." – I cannot get those numbers to correspond to what is hown in Fig. 2 – there I can see SMI for 33 males, and only 31 females. Why do numbers differ?

>**RESPONSE 48:** We apologize for the confusion. Because the purpose of creating the SMI categories was to look for a non-linear trend in SMI in relation to reproductive success, the sample includes only the grackles for which we have SMI and reproductive success data. That is 31 females and 33 males. We updated the Fig. 2 caption to reflect this detail (changes indicated in bold font):

"Frequency histogram of the SMI scores**, illustrating the SMI categories,** for the **33** males and **31** females **for which we also had reproductive success data**".

COMMENT 49: Repeatability of wing length: is this for repeatbale measures by the same person, or by different persons?

>**RESPONSE 49:** We measured the repeatability of wing length measures on the same grackle across time in 2 sampling sessions. And these second wing length measures could have been collected by the same experimenter, or a different experimenter. 11 out of 17 females and 10 out of 18 males were measured by the same experimenter on both occasions. We added this information to the new "Summary of methods" section after the Introduction:

"We tested the repeatability of our structural size measures on this subset of individuals by measuring them again before release. The second measures were collected by the same experimenter in 11 out of 17 females and 10 out of the 18 males that were repeatedly sampled."

COMMENT 50: You get a value of 0.53 – but what does that mean? Is this a low, moderate or high value?

>**RESPONSE 50:** The rpt function we use to calculate the repeatability value uses parametric bootstrapping, likelihood ratio and permutation tests to compare the observed repeatability value to a null hypothesis of no repeatability. Therefore, our observed value is significantly repeatable (From the manuscript: "Data permutations and a likelihood ratio test both confirmed that these repeatability values were statistically significant at $p < 0.01$").

Interpreting the repeatability value as low, moderate or high seems to depend on the question, study system and the experimenter's *a priori* expectation. For example, in Phillips & Furness 1998 values of 0.73, 0.45, and 0.26 are defined as high, moderate and low repeatability respectively, while in Rockwell et al. 2013 high repeatability is defined as greater than 0.5 and moderate repeatability as greater than 0.25.

Phillips, R.A. and Furness, R.W., 1998. Repeatability of breeding parameters in Arctic skuas. *Journal of Avian Biology*, pp.190-196.

Rockwell, C., Gabriel, P.O. and Black, J.M., 2013. Foraging dynamics in Steller's jays: Size and viability of cacheable food items. *Animal Behaviour*, *86*(4), pp.783-789.

COMMENT 51: Fat score: "Only 2 males were measured during the breeding season" – please state clearly in the methods that fat score values were from both breeding and non-breeding season. And motivate the use of values from the non-breeding season – I would like to know what a fat score in the non-breeding season may say about the individuals state in the breeding season and how this could be related to a males ability to attract females to his territory.

>**RESPONSE 51:** We created a new summary methods section after the introduction where we address your concern by stating:

"We also tested whether SMI and fat score varied by season because grackles are difficult to catch such that we were unable to structure our data collection to coincide with the breeding season and instead caught and measured grackles as often as possible."

Furthermore, fat score and SMI are likely important year-round for males because they also engage in dominance displays in the non-breeding season (Johnson & Peer 2001).

COMMENT 52: Table 1: what are the sample sizes for F and M?

>**RESPONSE 52:** We had fat score and SMI data for 47 females and 21 males. We updated the caption for Table 1 (now Table 2) to read: "Results from the logistic mixed-effect regression for ****47**** females and fixed-effect regression for ****21**** males…"

COMMENT 53: P2 – sample size: 20 females and 20 males is small, especially given your power analyses that calculates that you need a total sample size of 88 for a 90% chance to detect a medium effect.

>**RESPONSE 53:** Please see Response 34.

COMMENT 54: Fig 2: are the SMI scores presented here those for all F and M that could be measured, or only for those for which you also had data on reproductive success and hence could be included in analyses?

>**RESPONSE 54:** Please see Response 48.

COMMENT 55: Parts of the results seem focused on p values (p values displayed in text but no other stats, or p values that are also – and more informative with more other statistics – presented in tables). I would recommend making more use of effect sizes and their uncertainties and relate those to biological significance, instead of only displaying p values for which is not clear what they really tell.

>**RESPONSE 55:** Thank you for this feedback, we added more interpretation of the results based on the log odds effect sizes in the Results section as follows:

P1 - “In females, we found that for every one unit increase in SMI, the bird is 1.3 times more likely to have some fat (a 30% increase in the odds of having fat), which is not a statistically significant relationship (female *p* = 0.81; Table 2). In males, a one unit increase in SMI corresponds to an odds ratio of 1.6, or a 60% increase in the odds of having some fat, which is also not a statistically significant relationship (*p* = 0.50; Table 2). Together, this indicates that SMI and fat score are not equally measuring energetic condition. There was also no relationship between season (breeding or non-breeding) and female fat score (*p* = 0.71). Only 2 males were measured during the breeding season, therefore we omitted season as an independent variable in the male model (Table 1).”

P2, Effect of Season - “We used linear models to determine whether season would be important to include in our models **testing whether body condition relates to reproductive success**. We found that SMI did not differ by season **for females (Estimate (SE): *ß* = -0.30 (0.26), *p* = 0.26) or males (*ß* = -0.65 (0.43),*p* = 0.15). Similarly, fat score for females (*ß* = 0.28 (0.68), *p* = 0.68) and males (*ß* = 17.08 (2797.4), *p* = 0.99)** did not differ by season (Fig. 4). Although we note that, as stated above **and indicated in the standard error value**, we lack sufficient fat score data from males in the breeding season so **their** result should be interpreted with caution. Consequently, we did not include season as an independent variable in **our subsequent models testing the relationship between our body condition proxies and reproductive success**.”

P2, Main results - “Because fat score and SMI did not correlate, we included both as independent variables in our models testing prediction 2. For both males and females, we found no statistically significant relationships between either proxy of energetic condition and reproductive success (Table 3). **Of note, the inconsistent direction of the effects for the parameter estimates further supports that SMI and fat score do not measure the same trait.

For females, our SMI parameter estimate of -0.92 (exponentiated to get the log odds = 0.40) indicates that a one unit increase in SMI corresponded to a 60% decrease in the odds a female would fledge an offspring (*p* = 0.13). Whereas an increase from no visible fat to showing some fat corresponded to a 16% increase in the odds a female would fledge an offspring (log odds = 1.16, *p* = 0.82).** There was also no evidence of a significant relationship between the ability of a female to produce fledglings and having previously spent time in the aviaries **(log odds = 0.25, *p* = 0.22), where the odds that a female would fledge an offspring were 75% lower if females spent time in the aviaries.**

For males, there was also no statistically significant support for a relationship between whether a male defended a territory and SMI **(log odds = 3.25,** *p* = 0.13). **Nevertheless, this relationship may be biologically important because a one unit increase in SMI corresponded to a more than 300% increase in the odds a male will hold a territory containing nests.** Fat score was also statistically unrelated to male reproductive success **where an increase from showing no visible fat to showing some fat corresponded to a 28% decrease in territory holding (log odds = 0.72**, *p* = 0.76). Lastly, we found that those males who spent time in the aviaries were statistically less likely (**97% decrease in the odds**) to hold a territory compared with males who were never in the aviaries (**log odds = 0.03**, *p* = 0.02)..."

P2, logistic exposure - "We found that the probability of daily nest survival was significantly negatively related to SMI (**log odds = 0.50**, *p* = 0.03; Table 4), where, for every unit increase in SMI, the odds of daily nest survival decreased by half. This indicates that a female with a larger SMI (more mass for her structural body size) was less likely to have her nest survive each day (Fig. 5). There was no statistically significant relationship between the probability of daily nest survival and fat score (**log odds = 2.48**, *p* = 0.06), day of the year (**log odds = 0.81**, *p* = 0.16), or time spent in the aviaries (**log odds = 0.63**, *p* = 0.44, Table 4). Although not statistically significant, the effect size for the relationship between fat score and daily nest survival is large (Fig. 5) and potentially biologically meaningful. The odds of a nest surviving on a given day are almost 2.5 times greater (**248%**) for birds with some fat (a score of 1) compared to no fat (a score of 0)."

COMMENT 56: "spent more time in aviaries", "or who spent less time in the aviaries" -? According to Methods the aviary variable is either 0 or 1, but here it sounds as if you analysed how long time males spent in the aviary (and quite likely you have this data).

>**RESPONSE 56:** We apologize for this inconsistency. You are correct that the Aviary variable in our model was 0 or 1. We updated this sentence as follows:

"Additionally, we found that those males who spent time in the aviaries were less likely to hold a territory compared with males who were never in the aviaries (*p* = 0.02)."

COMMENT 57: "we did not have a balanced sample" – please give sample size for all possible combinations of aviary yes/no and hold territory with nests yes/no to show how balanced/unbalanced the sample was.

>**RESPONSE 57:** We created a new Table 1 with the sample sizes for the relevant variables.

COMMENT 58: Text below Fig 4: Here you stress that sample size was small. This you do when you present a result that was unexpected, while there is little stressing small sample sizes at other places where you do not find evidence for expected relationships.

>**RESPONSE 58:** We appreciate this comment as this bias is something that all authors should be careful of. We intended to caution the reader regarding the results where our sample size for males was small or unbalanced and therefore the results may not be as generalizable. In response to your comment 55, we added to the results and discussion more interpretation of the results based on effect sizes, in addition to p-values, for all of our investigated relationships. Please see Response 55 for specific changes to the text.

COMMENT 59: Table 2: what were the sample sizes for those analyses?

>**RESPONSE 59:** We added the sample sizes to the Table 2 (now Table 3) caption as follows:

"Results from the logistic regression for **20** females and **20** males…"

COMMENT 60: P2: what is the reason for relating daily nest survival only to female condition proxies? If parental provisioning is related to condition (state) and plays a role for nest survival then both F and M condition may have an effect? **Unless only females feed nestlings** – but I don't find any information about grackle parental provisioning in the methods.

>**RESPONSE 60:** You are correct that females do all of the nest building and parental care while the young are in the nest. We clarified in the 4th paragraph of the discussion as follows:

"We only included females in this analysis because males were never observed contributing to nest building, incubation, or feeding nestlings in our population and so will not have a direct effect on daily nest survival."

COMMENT 61: Reference to Table 3: should be Fig.

>**RESPONSE 61:** We are not sure what you mean in this comment. We double checked that all references to Table 3 (now Table 4) were appropriate, and could not find any that should have referred to a figure instead.

COMMENT 62: Disc: Another reason for lack of relationship may be that you measured in favourable years – mentioned in the Intro (effects of condition particularly in years with harsh env cond) but not discussed.

>**RESPONSE 62:** We found historical climate data for our study site location from the National Oceanic and Atmospheric Administration National Centers for Environmental Information (https://www.ncdc.noaa.gov/cdo-web/)

Table R1 below summarizes the average temperature and total precipitation recorded during the months of the breeding season (April - Aug) for three years prior to our study and the two years of our study (2019, 2020).

Table R1. Temperature and precipitation for Tempe, Arizona.

Year	**Min / Max Temperature (Fahrenheit)**	**Total Precipitation (inches)**
2016	66.9 / 98.4	2.31
2017	65.9 / 101	1.44
2018	67.9 / 100	1.41
2019	65.4 / 98.3	1.45
2020	69.4 / 101	0.5

So, the temperatures during our study were in line with those from the 3 previous years. Precipitation in 2019 resembled that in previous years. In 2020 there was less precipitation from April - Aug, but more precipitation earlier in the spring so it is unclear whether this should be designated as a more harsh year.

We commented on this in the discussion as follows:

"The increase in temperature variation and decrease in available water at our desert study site are both environmental stressors that have previously been found to negatively affect energetic condition [@pendlebury2004variation]. **Although our study spanned only two years, our data are likely representative of reproductive success in this environment because the temperatures during our study were in line with those from the previous three years [@noaa2020climate].**"

NOAA/NCDC, 2020: Climate Data Online Global Summary of the Month. NOAA/National Climatic Data Center. Subset used: April - August 2016–2020, accessed 7 July 2021.

COMMENT 63: Paragr two: from reading your text it seems that you equal energetic condition to stored fat. I am not a physiologist – so I wonder to what extent this may be true. What determines to short- and long-term energetic reserves of birds? Can they use other stores than lipids, e.g. proteins?

>**RESPONSE 63:** Research shows that stored fat is the primary source of energy in many taxa (Walsberg 1988), especially in birds (Pond 1981; Blem 1990). The energy per ounce from fat is much higher than from proteins or carbohydrates (Gessaman 1999). We clarified this in the discussion as follows:

"Second, SMI and fat score may measure different components of energetic condition because variation in mass among grackles could be attributable to muscle content, whereas fat score only accounts for subcutaneous fat [@labocha2012morphometric]. ****Research shows that stored fat is the primary source of energy in many taxa [@walsberg1988evaluation], especially in birds [@pond1981physiological; @blem1990avian] because the energy per ounce from fat is much higher than from proteins or carbohydrates [@gessaman1999evaluation].****"

Blem, C.R. 1990. Avian energy storage. In: Power, D.M. (ed), Current Ornithology. Vol. 7; New York; Plenum Press:59-113.

Gessaman, J.A. 1999. Evaluation of noninvasive methods of measuring avian body fat and lean mass. In Adams, N.J. & Slotow, R.H. (eds) Proceedings of the 22nd International Ornithological Congress: 2–16. Johannesburg: BirdLife South Africa.

Pond, C.M. 1981. Storage. In: Townsend, C.R. & Calow, P.(eds) Physiological ecology: an evolutionary approach to resource use. Sunderland, Massachusetts; Sinauer: 190-219.

Walsberg, G.E., 1988. Evaluation of a nondestructive method for determining fat stores in small birds and mammals. *Physiological Zoology*, *61*(2), pp.153-159.

COMMENT 64: Paragr two, "second": the statement that measuring muscle content requires sacrificing the birds is not true. Like using a mould to get an impression of the larger muscles, like breast muscle, that then can be use to quantify breast muscle size.

>**RESPONSE 64:** Thank you for correcting us here. We updated this part of the discussion to say:

"Measuring muscle content **often** requires destructive methods [i.e. sacrificing the birds; @zhang2015cross], ****or less objective assessments such as keel prominence or breast muscle shape [@gosler1991use; @abolins2017condition]**,** which was beyond the scope of the current research program."

COMMENT 65: Paragr two: Additionally there may also be an issue with lacking variation in your measures, at least for fat score. With little variation there is not much to show a relationship. But whether this may be another reason is hard to tell without any data on the distribution of values.

>**RESPONSE 65:** We added a new Table 1 with sample sizes, and we added the proportion of zero values for fat score in the Deviations from the planned methods section (see Response 47).

COMMENT 66: What do you mean by “larger females are more likely to disrupt nest stability”?

>**RESPONSE 66:** Thank you for pointing out that this statement lacked clarity. We’ve revised it in the discussion to say: “... or that larger females are unable to build nests in delicate vegetation structure that is more likely to be inaccessible to predators”

COMMENT 67: Related to this: what is the difference in size (wing length) and body mass between females with a low and females with a high SMI?

>**RESPONSE 67:** We are not quite sure we understand what you are asking here. Below is the mean and SE of wing length for females with low and high SMI, if we consider “low” SMI to be values less than the mean, and “high” SMI to be values greater than the mean:

Low: 138.4 (0.86) mm

High: 140.0 (0.89) mm

And the mean and SE for mass:

Low: 87.6 (0.93)

High: 100.0 (0.93)

COMMENT 68: I am confused with this study/manuscript. In the last paragraph of the Disc you outline what future research should do – but what you describe there is part of the validation (if done completely) that you seem to set out to do in the Intro. At least your Intro (aiming an validation of proxies) raises those expectations.

>**RESPONSE 68:** Please see Responses 20, 27, 29, and 30 above for how we address your concerns on proxy validation and the framing of our study design.

COMMENT 69: Relation condition – reproductive success: it would be useful with e.g. frequency of nest visits with food or other behavioural/performance traits that relate to reproductive success.

>**RESPONSE 69:** We received pre-study reviews and in-principal acceptance of the study plan in the current state, so we cannot change it now.

COMMENT 70: Methods: There is no information about:

-at which time of the year you caught grackles and took the measurements.

>**RESPONSE 70:** We caught grackles year-round. See Response 51 above for how we added this information.

COMMENT 71: -how measures (proxies) vary across time of year.

>**RESPONSE 71:** We compared SMI and fat score values in birds measured in the two distinct seasons that non-migratory birds experience (breeding and non-breeding). We found no evidence for variation in these proxies with season. (From the manuscript Results section: We found that neither SMI (female $p = 0.26$, male $p = 0.15$) nor fat score (female $p = 0.68$, male $p = 0.99$) differed by season in females or males (Fig. 4).)

COMMENT 72: -whether measurements of condition were only used when taken during the same year as there were data on reproductive success of the same individual, or if they were used when taken in anyone year.

>**RESPONSE 72:** It was logistically prohibitive to catch grackles multiple times to get measures of energetic condition in the months prior to the breeding season each year. We assume that because we found no difference in SMI or fat score by season that an adult bird's body condition

would not significantly vary across years when environmental variation is less dramatic than across seasons.

The reproductive success data from 2019 included females and males in which all energetic condition data were taken in the winter or early spring prior to the breeding season.

Of the 13 females with reproductive success data in 2020, 6 were measured in the winter or early spring prior to the breeding season and 7 were measured at an earlier date in 2018 or 2019. For males, 11 were measured in the winter or early spring prior to the 2020 breeding season and 6 were measured at an earlier date in 2018 or 2019.

Review by Tore Slagsvold, 2021-01-28 15:10

****COMMENT 73:**** General: The authors have studied two indices for body condition and relationships with reproductive success, in great-tailed grackles, a theme of great general interest. The study aims to be of general significance, at least to avian ecology. However, there are some problems.

>RESPONSE 73:**** We appreciate the time you dedicated to reviewing our manuscript. We address your concerns below. However, as this is a Stage 2 review of this registered report, we are not allowed to significantly change components of this manuscript (Introduction, Hypotheses, Methods and Analysis Plan) that have already been reviewed by multiple other reviewers and received in principle acceptance by PCI Ecology. We apologize for any inconvenience that this confusion about the Stage 2 review process may have caused you.

****COMMENT 74:**** (1) Sample size is low, as the authors admit. Thus, one may question whether the chosen species and the population studied is suitable for the question asked. For instance, with the small samples, it is almost impossible to obtain reliable results from analyses of condition indices and of linearity of the variables. The situation is even more complex when the authors include birds that have been kept in captivity. **Effect of captivity may be presented in a separate paper.** Thus, I recommend **more data to be collected**.

>RESPONSE 74:**** Please see Response 34 where we address concerns about sample size. Additionally, we explicitly tested for an effect of captivity and found that it had an effect on males but not females (Table 2 in the manuscript). If we excluded the effect of captivity from this paper, then there would be variation in our response that is unaccounted for by this variable known to affect life history. If we excluded captive birds, then our sample size would be even smaller.

COMMENT 75: (2) Reproductive success was measured as whether or not a nest fledged offspring. This is a quite coarse-grained measure compared to what is readily available for many other bird species. And it is critical when it is in combination with small sample sizes (data only for 20 females and 20 males).

>**RESPONSE 75:** We received pre-study reviews and in-principal acceptance of the study plan in the current state, so we cannot change it now.

COMMENT 76: In addition, apparently, **the measure did not take into account nest failures caused by predation, i.e. losses that are not necessarily related to female body condition**. Thus, the authors should add information as to which extent nest predation was involved.

>**RESPONSE 76:** We received pre-study reviews and in-principal acceptance of the study plan in the current state, so we cannot change it now.

Additionally, please see our Response 21 above.

COMMENT 77: Please add already in Results how many nests that failed. In studies of condition indices,

>**RESPONSE 77:** We added a new Table 1 with detailed sample sizes for the relevant variables including a *Prop. successful* category that represents the proportion of the total individuals observed engaging in breeding behaviors in each year that held a territory containing nests (males) or fledged young (females).

COMMENT 78:I would have preferred to include only nests where at least one chick fledged in the analysis.

>**RESPONSE 78:** We received pre-study reviews and in-principal acceptance of the study plan in the current state, so we cannot change it now.

COMMENT 79: (3) In birds, great variation exists in fat score and body mass not only among individuals and years, but with time of season and time of day. This has to be taken into account in any study of the relevance of condition indices. However, I do not find a mentioning of this. **In the result-section, you mention the variable “Season”, however, you do not seem to define it**. You say that only two males were measured in the breeding season: I wonder at which stage of the breeding season, and in particular, when were all the other birds sampled.

>**RESPONSE 79:** Please see Responses 25 and 44 where we address concerns about temporal variation in energetic condition proxies.

Please see Methods > Independent Variables > P1 in the manuscript for our definition of Season.

COMMENT 80: (4) In the introduction, there are many examples from ectotherm animals (vipers, mantids, crickets). However, I do not think such species are relevant here when the present study deals with birds. Birds are endotherms, and in addition they are selected to keep low body mass to save flight costs and escape predation.

>**RESPONSE 80:** Thank you for bringing this up. Please see Response 85.

COMMENT 81: (5) End of introduction. You say that your model birds are ideal for such a study because they face severe physiological and ecological challenges. You may support this statement with data on survival rates and their variation among seasons.

>**RESPONSE 81:** We received pre-study reviews and in-principal acceptance of the introduction in the current state, so we cannot change it now. However, our study spanned 2 years and thus most color-marked individuals were seen throughout this period and those that disappeared did so for unknown reasons. Additionally, see Response 2 above for how we elaborated on the suitability of this species for this question in our discussion section.

COMMENT 82: (6) You analyzed repeatability of the wing length measurement. How long time had elapsed between the two events. Did you test for repeatability across different persons? You say that it may have been difficult to assess the fat score. Did you test for repeatability?

>**RESPONSE 82:** The only grackles that we measure twice are those that are caught to be brought into aviaries, then caught again within the aviaries to be released back to the wild. The time between measures depended on how quickly birds completed the behavioral tests in the aviaries,

which ranged from 1 to 6 months apart. Because birds had more access to nutritious, supplemented food and water while in the aviaries, the second measure of fat score could be artificially increased, therefore fat score is unlikely to be repeatable. Our repeatability analysis tests whether we consistently measure wing length across time over the same and different experimenters. These details are currently included in the new Summary of Methods section after the Introduction and the Analysis Plan section towards the end of the article.

COMMENT 83: (7) It is fine that authors present the relevant hypotheses, and the predictions, early in a paper. However, in this case, with all the problems mentioned above, I wonder whether it is worth it because it takes space, and the points and arguments have to be repeated in any case in the Discussion. Perhaps it is be better simply to tell which indices you will study, and then go directly on the results, and finally discuss the results in more detail in the Discussion chapter.

>**RESPONSE 83:** We received pre-study reviews and in-principal acceptance of the hypotheses and predictions in the current state, so we cannot change it now.

Review by Kyle Elliott, 2021-01-09 04:19

COMMENT 84: The use of scaled mass index (SMI) as a measure of 'body condition' is being adopted widely with little critical review of this metric. SMI purports to improve on early metrics, such as the residual of mass on size, in two ways: (1) by using a log-log approach given the allometry between linear size and mass (~size ^ 3) and (2) using reduced major axis regression because there is error in both measurement of size and mass. The authors examine the relationship between SMI and lipid scores in breeding grackles and find no such association, suggesting that SMI is not a good index of 'body condition' in those birds during the breeding season. They also find no relationship between lipid stores or SMI and breeding success, suggesting that they are not good indices of fitness. The paper was pre-registered, which I appreciate, and the statistical analyses are carried out carefully. Moreover, given how widespread SMI has become, papers such as these, which challenge those assumptions, are critically important. They also have a very nice discussion of some of the issues with naive use of SMI and similar condition indices.

>**RESPONSE 84:** Thank you very much for reviewing our manuscript. In particular, we are grateful for the relevant papers you have brought to our attention. Your comments and suggestions have helped to vastly improve this manuscript.

COMMENT 85: However, I do have several points that I think need to be acknowledged in the Discussion. Much of the Introduction is focused on the non-avian literature. I appreciate that the authors are attempting to broaden their paper's relevance, but I think there are important references from the avian/bat literature that could be incorporated. Body mass is clearly especially important for flying animals where costs are closely related to body mass (takeoff speed compared to predators, energetic cost of staying aloft), so I think it is especially worth examining the literature in birds and bats. Both papers below examined SMI vs. lipids in multiple bird/bat species and found no relationship, supporting what you found.

McGuire, L.P., Kelly, L.A., Baloun, D.E., Boyle, W.A., Cheng, T.L., Clerc, J., Fuller, N.W., Gerson, A.R., Jonasson, K.A., Rogers, E.J. and Sommers, A.S., 2018. Common condition indices are no more effective than body mass for estimating fat stores in insectivorous 1debats. Journal of Mammalogy, 99(5), pp.1065-1071.

Jacobs, S.R., Elliott, K., Guigueno, M.F., Gaston, A.J., Redman, P., Speakman, J.R. and Weber, J.M., 2012. Determining seabird body condition using nonlethal measures. Physiological and Biochemical Zoology, 85(1), pp.85-95. (I am a co-author on one of those papers, as well as Nip et al below, and feel free to replace with other papers; my suggestion is to acknowledge that SMI has not had a lot of support across many studies that have addressed SMI vs. lipids in birds)

Gosler, A.G., Greenwood, J.J. and Perrins, C., 1995. Predation risk and the cost of being fat. Nature, 377(6550), pp.621-623. (not on SMI, but I believe Gosler was one of the first to really emphasize the cost associated with being heavy, rather than McNamara, so might be a relevant reference to add).

>RESPONSE 85:**** Thank you very much for these suggestions! We incorporated these citations into the introduction and more content on birds and bats in the discussion as follows:

In the introduction: "...A similar lack of a relationship was found in flying animals such as birds [@gosler1995cost; @jacobs2012determining] and bats [@mcguire2018common] thus indicating cross-taxa support that morphological proxy measures do not always measure the same trait. "

In the discussion we added: "...we found that two proxies of energetic condition, fat score and SMI, did not correlate with each other in the great-tailed grackle, regardless of whether it was the breeding or non-breeding season. ****This has also been found in studies on bats [@mcguire2018common], which are species that similarly experience distinct demands on body structure to facilitate flight.****"

****COMMENT 86:**** You rightly challenge SMI/lipid stores as a measure of condition, but body mass change in breeding is more nuanced than you present. For one thing, change in body mass (which would presumably be the same thing as SMI if they authors bothered to convert into that) in breeding bird is often thought to be 'adaptive'. That is, birds lose mass at chick-rearing when they make many trips to feed offspring so as to reduce flight costs (Freed 1981; Gaston and Perin 1993). **It's unclear**

to me that you have carefully selected birds at the same stage of breeding. If not, then one would not expect any relationship between body mass (or SMI or lipid stores) and fitness. I could imagine that body mass (or SMI or lipid stores) of females at the onset of breeding could predict subsequent success, but that those individuals with more offspring might subsequently have more mass loss, so lumping everything together would lose any such relationship. It seems important to acknowledge that mass (or SMI or lipid stores) might be 'programmed' to drop over the breeding season. I think body mass change through the season is an important mechanism to add to the third paragraph of the Discussion.

Freed, L.A., 1981. Loss of mass in breeding wrens: Stress or adaptation?. Ecology, 62(5), pp.1179-1186. Gaston, A.J. and Perin, S., 1993. Loss of mass in breeding Brunnich's Guillemots Uria lomvia is triggered by hatching. Ibis, 135(4), pp.472-475.

>RESPONSE 86:**** Per our permit regulations, we did not catch any grackles that we knew were actively building, incubating, or feeding nestlings so we should not see variation among females due to measurements occurring at different stages of the breeding process. However, you are correct that, due to logistical constraints and the difficulty of catching grackles, we were unable to carefully select birds to measure at the same time of year (or time of day). Although we include measures from grackles caught in the breeding and non-breeding months, we did not find a significant effect of season on the energetic condition variables.

We added the following to the discussion on this point:

"Although we found no evidence that SMI or fat score varied by season, there is evidence from other studies that avian mass changes with time of day [@nip2018seasonal] and stage of breeding [@milenkaya2013variation]. It was logistically impossible in our project (and in many avian research programs) to capture birds multiple times within a season or several times per day, therefore temporal variation in data collection could obscure the correlation between these two proxies. However, the stage of breeding is unlikely to introduce additional variance to our study because we did not catch any females that were actively engaged in any stage of the breeding process."

****COMMENT 87:**** I think it would be worth separating 'body condition' which is used so widely as to be useless from size-corrected mass which is a specific idea that may or may not be related to anything else.

>RESPONSE 87:**** See Response 36 above where we change all "body condition" or "condition" terms to "energetic condition" to clarify.

COMMENT 88: The authors mention the ratio methods ("ratio of weight to tarsus length") but not the residual methods which are more widely used (and, for instance, correlated better with lipid stores than SMI in Jacobs et al. (2012).

>**RESPONSE 88:** Thank you for this feedback, we updated the wording in our introduction to read:

“A variety of morphological proxies have been used to quantify energetic condition [i.e., fat score, weight, **ratio of mass to structural size, residuals from a linear regression of mass as a function of structural body size; @labocha2014body; @jacobs2012determining]**”

COMMENT 89: Breeding songbirds, especially desert birds, may not be the best test of SMI vs lipid scores. As you point out, your lipid scores were uniformly low (0 or 1). Jacobs et al. (2012) points out that condition indices work best when % lipid is high and there is enough variation to examine. For example, wintering chickadees that show large variation in lipid scores show a correlation between SMI and lipid score (Nip et al. 2019). Given that 70% of a bird is water, and desert birds might become particularly depleted in water, I could imagine that any variation in body water might overwhelm variation in lipid mass (rather than just muscle mass as presented in the Discussion). It would be helpful to note that your results fit in with the prediction from Jacobs et al. that birds with low percent lipids have weak relationships between condition indices and lipids. This would be an additional explanation to bring up in the second paragraph of the Discussion.

Nip, E.J., Frei, B. and Elliott, K.H., 2019. Seasonal and temporal variation in scaled mass index of Black-capped Chickadees (Poecile atricapillus). The Canadian Field-Naturalist, 132(4), pp.368-377.

>**RESPONSE 89:** Good point, we added this to the discussion as follows:

“...because variation in mass among grackles could be attributable to muscle **or body water content… However, because desert birds have inconsistent access to water sources, variation in body water content may obscure variation in lipid content.** ”

COMMENT 90: Wing and tarsus are probably not the best measures of structural size. Wing length could be related to conditions in previous moult rather than determinate growth. Tarsus is notoriously unreliable as a metric, due to difficulty in repeatability of measurement. Wing also has low repeatability (see Hull et al.) A better size metric is possible head + bill.

Hull, C.L., Vanderkist, B.A., Lougheed, L.W., Kaiser, G.W. and Cooke, F., 2001. Morphometric Variation in Marbled Murrelets, Brachyramphus Marmoratus, in British Columbia. Northwestern Naturalist, pp.41-51.

>**RESPONSE 90:** We appreciate the feedback on these measures. In the grackles that we hold in aviaries for behavioral testing, we've noticed obvious bill growth that is potentially related to lack of substrate variety that normally abrades the bill tip. Therefore, we thought this was likely a poor measure for structural body size in our system.

COMMENT 91: The authors describe SMI in some detail twice. I think you only need to do so once, and **given that its widely used, I'm not sure that more than one sentence giving the equation is needed**. I think you can also challenge the SMI theory. Peig & Green argue that RMA regression is better because mass and length both have error. However, when you are actually trying to predict one of the variables (i.e. mass) then OLS is a more accurate representation.

>**RESPONSE 91:** Good point, we removed the SMI equation from the methods section. We added to the discussion the following to address your point on RMA vs OLS:

"Because SMI can perform poorly in birds with low lipid mass, future research should also compare several mass by structural body size equations to determine the most appropriate proxy for a specific study system [@jacobs2012determining]."

COMMENT 92: Figure 3. It seems strange that the relationship with males is not significant.

>**RESPONSE 92:** Yes, it does seem like there is a potential relationship here that we were maybe unable to detect because of low sample size. Per comment 55 above, we added an additional interpretation of the parameter estimates to our results, so readers can form their own opinion on whether a relationship may be detected with additional data:

"For males, there was also no statistically significant support for a relationship between whether a male defended a territory and SMI (log odds = 3.25, *p* = 0.13). Nevertheless, this relationship may be biologically important because a one unit increase in SMI corresponded to a more than 300% increase in the odds a male will hold a territory containing nests."

And we also noted in the discussion:

"Although our results were not statistically significant, in some cases the parameter estimates revealed log-odds that may be large enough to be biologically significant. Notably, a one unit increase in SMI corresponded to a more than 300% increase in the odds a male will hold a territory containing nests, but a 60% decrease in the odds a female would fledge an offspring."

COMMENT 93: "Future research could add to this work by incorporating additional methods to measure energetic condition, for example, blood hematocrit levels (Dawson and Bortolotti 1997), protein storage (Houston et al. 1995)". Those are almost certainly very poor metrics of 'condition' (see review by Fair which found no support Hct, except for birds parasitized by blood parasites). I suggest (1) accounting for stage of breeding (day since laying of first egg), and (2) measuring total lean mass or total lipid mass using quantitative magnetic resonance.

>**RESPONSE 93:** Thank you for this feedback; we have revised the discussion to reflect these comments. We incorporated your suggested HCT citation and QMR suggestions as follows (changes in bold):

"**If financially and logistically feasible, future research could measure total body composition and relative mass of fat using the relatively new and promising method of quantitative magnetic resonance [@guglielmo2011simple],** or researchers could incorporate additional physiological methods to measure energetic condition, for example, blood hematocrit levels [@dawson1997avian; **but see @fair2007sources**]."

We incorporated additional comments about accounting for stage of breeding (or temporal variation, generally) as follows:

"Although we found no evidence that SMI or fat score varied by season, there is evidence from other studies that avian mass changes with time of day [@nip2018seasonal] and stage of breeding [@milenkaya2013variation]. It was logistically impossible in our project (and in many avian research programs) to capture birds multiple times within a season or several times of day, but temporal variation could obscure the correlation between these two proxies"

COMMENT 94: Line numbers would have been helpful.

>**RESPONSE 94:** We apologize for this inconvenience and oversight. You are absolutely right we should have included line numbers in the pdf.