THE ROYAL SOCIETY PUBLISHING

PROCEEDINGS B

Longer telomeres during early-life predict higher lifetime reproductive success in females but not males

Britt J. Heidinger, Aurelia C. Kucera, Jeff D. Kittilson and David F. Westneat

Article citation details
Proc. R. Soc. B **288**: 20210560.
http://dx.doi.org/10.1098/rspb.2021.0560

Review timeline

Original submission:	30 December 2020
1st revised submission:	9 March 2021
2nd revised submission:	21 April 2021
Final acceptance:	4 May 2021

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2020-3219.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Acceptable

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
Yes

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
No

Is it clear?
N/A

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors report that among-individual TL variation of wild house sparrow during early life positively predicts lifetime reproductive success, and that this effect was sex-specific (it emerged for males but not for females) and entirely mediated by differences in longevity. I think the ms is generally well-written and clear. I do have however some major concerns (labeled M), as I will detail below, that I urge the authors to carefully consider in order to make their case stronger. All comments are written in order of appearance in the ms.

L1: I found the title rather misleading upon carefully reading the ms. Actually, if longer telomeres non-significantly predict reproductive success upon controlling for longevity, stating 'predicts higher lifetime fitness' can lead to the false impression that TL predicts reproductive success. I would thus rephrase it as follows: "....predicts greater female but not male longevity"

L33: the reference to 'reproductive rates' is somewhat obscure here and can be understood only upon reading carefully the results. I suggest rephrasing this sentence to make it clearer.

L75: telomeres being a correlated

L76-77: suffer lower reproductive costs

L81: few non-human studies have

M L107 and subsequent: far more details are needed here about the study population characteristics and how you have calculated the fitness endpoints in this initial paragraph about the study system, as detailed below.

First of all, it is unclear whether this population shows breeding dispersal. The following information must be provided in order to make a compelling argument that breeding dispersal does not affect your conclusions: 1) what was the age at first breeding of all individuals included in the analyses of longevity and breeding success? 2) what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?

You should also made clear that you are focusing on a very specific sample of individuals, i.e. local recruits, which represent a tiny fraction of the total fledglings from this population. I am not saying this can introduce bias, but we know that these individuals may be somewhat odd, as the vast majority show natal dispersal, and there is evidence from other species that local recruits

may differ from dispersing recruits in many different ways, especially in species where local recruitment is very low; they are extreme individuals, so it is unclear how patterns detected in these individuals reflect those occurring in the population at large. A word of caution should be added. Moreover, is natal dispersal/local recruitment sex biased? In many passerines, recruits are mostly males. What is the sex ratio of your local recruit population? Did you compare longevity of local recruits (males, females) with that of immigrant individuals (for which you should have data)?

Clearly, providing this information is essential to convince the reader that you have measured both longevity and LRT in a robust way.

Finally, you need to report in this section in detail how you have calculated fitness components (longevity and LRT). These variables are only first mentioned in the statistical analysis section, while they should be introduced and described here.

L175: I am not sure whether I would include plate ID in models as a fixed effect. Basically, it is a nuisance variable you aim at controlling for, you do not wish to test specific hypotheses about this variable. Why not including it as a random effect instead? An easier alternative would also be to remove plate effects by within-group centering (i.e. standardizing plate values to the reference sample value, mentioned in L154).

L182: I guess you meant 'a by-subject random slope with age' here, right? I am skeptical it is feasible to model a random slope age effect at the individual level when you have a maximum of 4 datapoints per subject (if I understood correctly). Please check this out. Nevertheless, this model shows that, despite the population shows a (non-significant!) tendency for a decrease in TL with age, individuals significantly differ in their age-related variation in TL, with some increasing and others decreasing. This individual variation and its implications for this study (if any) is not mentioned in the results.

L185: rephrase: ...freedom estimated according to the Kenward-Rogers method, and random effects...

L192: this is confusing. Of course if you had local recruits they must have bred at least once. Using the 'local recruit' definition would make everything clear. You should rather state here that birds were blood-sampled at least once upon settling as breeders in the population.

M L195 and subsequent. You should made it clear that you have verified a key assumption of Cox regression, i.e. a constant multiplicative effect in the hazard function over time for each covariate. Currently, this cannot be assessed (not even visually) as you are nowhere showing a survival curve. I urge you to present survival curves (e.g. in relation to sex) and to explicitly verify that Cox proportional hazard model assumptions are met. Otherwise, you can consider running a GLM with longevity as a dependent variable and sex and TL10 as predictors. This is the way you are showing it in Fig. 2A, by the way....of course in that case you should take care of the distribution of longevity (likely not purely Gaussian).

L198-205: move in first subheading of methods (see previous comment). Also, specify how you could estimate brood size (at which age? how much before fledging?). More details are needed here to properly assess the robustness of your results.

L211-212: the standard term for this trait is 'laying date'. Please refer to laying date throughout (instead of date of first egg, etc.).

L213: please refer to 'dataset', not 'datafile' (here and elsewhere).

L214-215 and subsequent: very poorly presented and confusing. What do you exactly mean by 'included some known covariates'? Please report clearly which covariates and which interactions

were included in each model.

L216-217: year should be included as a random rather than fixed effect. You are not interesting in testing hypotheses about year differences.

L221-222: 'assuming' is not the best word here. You can test these 'assumptions' and check which distribution fits best your data. Sounds weird that Clutch size is Gaussian and fledging success is negbin. Please check it out. Then, once more, be more coherent with naming your variables, because it seems you are confusing fledging success with number of offspring. I agree that modelling the number of offspring with clutch size as an offset in practice equals to analyzing fledging success, but it is not correct to say that you analysed fledging success. Moreover, the standard term for 'number of offspring' is brood size. Please use this term throughout.

L231-232: actually, this is unclear to me because this formula does not refer to a 'bivariate' LMM, where TL and fitness are both dependent variables. What I see is that the two LMM (for TL and for LRT/longevity) here have different formulas, and I am unsure whether this is ok. Overall, I'm unfamiliar with BLMM, so I admit I may be wrong here. However, I am aware that the general recommendation is to avoid relying on BLMM unless you have a very large datasets with many repeated observations per individuals, which should be preferably collected at the same time point (not the case here, right?). I would thus avoid using this approach and remove these analyses from the manuscript (although they might be moved to the suppl. mat). Indeed, on the one hand you manage to avoid performing stats on stats, but on the other hand BLMM may lead to spurious conclusions as assumptions may be easily violated and power is generally low unless very large samples are available (see Dingenmanse and Dochtermann JAE). I would personally trust more results deriving from 'stats on stats' in this case...

L271: although significant, it should be specified that repeatability is definitely low.

L274: this statement is questionable given 1) the scatter shown in Fig. 1 and 2) the p-value of 0.053 for the age effect.

L275-276: this way of presenting results is confusing, in my opinion. Here you are testing the effect of sex, not F-M. Rather, specify in the methods that you have coded sex as 0 = female and 1 = male (correct?). What does F-M means? Difference in estimate between female and male values?

L278-280: be careful because the correlation of estimated slopes and intercepts may not have any actual biological value (https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#singular-models-random-effect-variances-estimated-as-zero-or-correlations-estimated-as---1) as it may depend on lack of convergence or other issues with model specifications. I recall reading some comments by mixed model 'gurus' stating that high correlation generally reflect model mis-specification and collinearity among predictors (which may considerably change by using e.g. centered predictors). I would personally avoid mentioning or interpreting these stats.

L283-284: this was puzzling to me. How could you estimate the change in TL with age for birds that did not breed? Were they captured during the non-breeding season when birds from nearby breeding sites gather together while moving to breeding sites which were not monitored? This requires some additional explanation in the Methods.

L292-293: again, a confusing way of presenting results from statistical analysis. Please be explicit in stating that you are testing an interaction term here (sex x TL10 on longevity).

L308: less than in females

L333-334: I find confusing here referring to 'lifetime success' and including longevity. I think you should always better separate LRS from longevity.

L333-335: where is the evidence for lack of sex differences? Please add reference to Table 3.

L339: again, a confusing mention of 'lifetime fitness' here. See previous comment. Be consistent with terminology throughout.

L411: wording unclear, rephrase: ....across species showing disparate life histories....groups remains to be elucidated.

L416: correct typo

Fig. 1: clarify what 'lines' are here. It seems these are 'regression lines' for each individual. I would rather show lines connecting individual datapoints for a given individual (it does not make much sense to compute slopes based on 4 datapoints...how could you assume the relationship is 'linear'?). BTW, this figure makes me wonder whether it would be worth using TL10 as a log-transformed variable, as it seems it is highly skewed. Also including it as a predictor may lead extreme values to have a high leverage on regression/mixed model output. I am unsure whether you did test it, but I would consider this carefully.

Fig. 2: confusing, as I do not think presenting separate 'linear regressions' is appropriate here. Moreover, I guess reference to GLMM here is wrong, as you have a single datapoint per individual, right? (see L 207-209). Please correct to GLM. And you should include confidence bands for each group (can be easily obtained via the 'visreg' R package). The best would be to plot 'raw' data with GLM-estimated slopes by group.

Table 1: statistics not properly reported. Under the random effects, Intercept should read 'Individual identity'. I also believe that 'plate' should be a further random intercept effect (or that it should be treated in a different way, as I suggested).

Table 2: confusing, requires improvement. Male-female should be Sex. Interactions are reported in a confusing way. Moreover, if year is included as a fixed effect (which I do not agree, by the way...why not as a random effect?), the statistics (global F test) should be reported. Why 'Subject age'? Never heard of this variable. Do you mean 'Age'? What do you mean by Telomere length at day 10 (Female)? And by Telomere at day 10 (M-F)? Also, list all predictors in the same order in all sub-tables, at present they are confusing. Why 'Age' is not included in models of clutch size and Offspring? To sum up, you need to present these results in an inconsistent and confusing manner which needs to be improved. Factor by covariate interactions should be reported as e.g. Sex ×TL10, with global F-test and associated estimates (one for each sex category), to highlight how slopes diverge. Check out the excellent 'reghelper' R package, command 'simple_slopes' to get group-specific slopes easily without too many calculations and recoding. The 'usual' way of reporting differences from a reference value is misleading and unclear.

Table 2: confusing, requires improvement. See previous comments about 'lifetime success' and how it generates confusion. The meaning of the variables (Females and Male-Female) are unclear. Also, chi square tests are unclear here: what do they represent? I would expect to see one slope for males and one for females. Where do these statistics come from? Which kind of models? Which variables are included? These are not properly described in the methods. Are models ran separately for each sex? What is the sample size per sex? These details must be provided.

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
Yes

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is a well written, carefully thought-out, and comprehensively conducted analysis. The fact that the analysis script and data are included is a further laudable strength. I also appreciate the model robustness checks with supplemental Bayesian models (although I did not evaluate the Bayesian models). This should be published after minor revision. My specific suggestions below are all pretty minor points.

Title: if there is space would be good to fit in the species name

L53 – TLs shorten with age and predict longevity in SOME species—both of these patterns appear to be substantially heterogeneous across species

L62 – While this sentence is justified based on the broader literature—the findings of Epel et al have not held up to replication, and I would suggest citing a meta-analysis or seminal finding which does hold up. Same point applies to line 65—while a few studies do show reproduction

predicting shorter TL in women, the Epel study is looking at caregiving for disabled children, which is not the same thing—and has not replicated well

L68 – TL loss suggests longitudinal change was measured—which was not done in all of these studies

Telomere length measurement methods are generally very well described and in much more detail than is typically provided. Nonetheless, please be sure all portions of the emerging best practices from the telomere research network are considered: https://osf.io/9pzst/

In particular, T/S ratios should be transformed to z-scores to make effects more interpretable following the TRN and Verhulst 2020 suggestions.

L134 - While GAPDH only having one amplicon size as discerned by melt curve and electrophoresis is a good check, it is still possible that the amplicon varies in copy number across individuals or has genetic polymorphisms which influence amplification efficiency. Such patterns would also tend to increase intra-individual repeatability of T/S ratios. This could be addressed by re-running samples with the highest and lowest GAPDH concentrations and trying another putative single-copy gene control to be sure it and the GAPDH results are similar. I don't think this experiment should be a requirement for publication, but would help to assure that TL is measured well here.

L152 – this is fine and reasonable. I would suggest that the authors consider using methods to adjust for efficiencies that deviate from the ideal of 2 (aka 100%) in their future work—but do not think this should be required for publication. While efficiencies were measured, they do not seem to have been used for calculating T/S

L158 – what were R^2 values of standard curve?

L161 – good—but please be more specific about which version of the ICC statistic was calculated here—there are several

Controlling for baseline TL can confound examination of longitudinal changes in TL (Bateson et al. 2019). Please double check that this error is not being made in your analyses.

L244 – I am a little concerned that ln transformation is changing the nature of the hypothesis being tested in a subtle way—that is you are now testing how TL predicts ln number of eggs and ln longevity which is a non-linear and biologically different hypothesis.

Figure 1 – nice and helpful figure—I suggest the authors consider making the lines and dots narrower/smaller, to allow readers to better discern the patterns

L278 - Greater variance in TL change in females is an interesting finding. Can you put a p value to this difference between sexes and/or show confidence intervals of these estimates so we can get a better sense for how different these are when considering the uncertainty of the estimates? Also, I don't think this finding was further considered in the discussion and probably should be.

L293 – would be easier to read if terminology was kept more consistent across the text and figures. It appears longevity, probability of disappearing and cumulative survival all have identical meanings—or if there is a distinction being made between these, it is not clear to me.

L329 – please make more clear in the text here the nature of the collinearity? i.e. does longer baseline TL predict greater decline in TL consistent with regression to the mean expectations?

It occurs to me that if it is blood TL which is having the effects on biology than blood TL measured later in life should be more predictive of reproductive success and longevity AFTER

that TL measure than the earliest TL measure. I don't think this was directly tested in this paper. Given how comprehensive the analysis already is, I think the authors should feel free to test this or not at their own discretion.

Bateson M, Eisenberg DTA, and Nettle D. 2019. Controlling for baseline telomere length biases estimates of the rate of telomere attrition. Royal Society Open Science 6(10):190937.

Decision letter (RSPB-2020-3219.R0)

01-Mar-2021

Dear Dr Heidinger:

I am writing to inform you that your manuscript RSPB-2020-3219 entitled "Longer telomeres during early life predicts higher lifetime fitness in females but not males" has, in its current form, been rejected for publication in Proceedings B.

This action has been taken on the advice of referees, who have recommended that substantial revisions are necessary. With this in mind we would be happy to consider a resubmission, provided the comments of the referees are fully addressed. However please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

1) A 'response to referees' document including details of how you have responded to the comments, and the adjustments you have made.
2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
3) Line numbers in your main document.
4) Data - please see our policies on data sharing to ensure that you are complying (https://royalsociety.org/journals/authors/author-guidelines/#data).

To upload a resubmitted manuscript, log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,
Dr Locke Rowe
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
Your manuscript has now been reviewed by two experts in the field and myself. We all agreed that the question is interesting, taking advantage of a long term dataset in house sparrows, and that the manuscript was well written. As such, there is good potential for this manuscript to make a valuable contribution to our understanding of how telomere dynamics relate to individual quality. There were, however, several concerns raised by both reviewers that would need to be addressed before further consideration is possible.

Both reviewers mention concerns with the analyses, with reviewer 1 providing several detailed comments that range from how variables are being categorized in the models, to the validity of some of the current statistical approaches. Reviewer 2 raised concerns with the T/S ratios not being transformed to z-scores and questioned the use of ln transformation and how that might affect the ability to test the stated hypothesis. They also commented on some of the terminology being used and whether it was consistent within the manuscript or with terminology used by others.

Reviewer 1 requests additional information on the characteristics of the study population, pointing out that individuals that do not disperse from their natal habitat might not be good representatives of the whole population. They also suggest that a discussion of the measured fitness components should come earlier in the manuscript.

Reviewer 2 cautions against the use of the Epel citation and mentions several potential concerns with the qPCR methods. Even if not addressed through additional analyses, a more thorough discussion in the methods may be warranted.

I concur with these concerns and believe that addressing them will strengthen the manuscript.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
The authors report that among-individual TL variation of wild house sparrow during early life positively predicts lifetime reproductive success, and that this effect was sex-specific (it emerged for males but not for females) and entirely mediated by differences in longevity. I think the ms is generally well-written and clear. I do have however some major concerns (labeled M), as I will detail below, that I urge the authors to carefully consider in order to make their case stronger. All comments are written in order of appearance in the ms.

L1: I found the title rather misleading upon carefully reading the ms. Actually, if longer telomeres non-significantly predict reproductive success upon controlling for longevity, stating 'predicts higher lifetime fitness' can lead to the false impression that TL predicts reproductive success. I would thus rephrase it as follows: "....predicts greater female but not male longevity"

L33: the reference to 'reproductive rates' is somewhat obscure here and can be understood only upon reading carefully the results. I suggest rephrasing this sentence to make it clearer.

L75: telomeres being a correlated

L76-77: suffer lower reproductive costs

L81: few non-human studies have

M L107 and subsequent: far more details are needed here about the study population characteristics and how you have calculated the fitness endpoints in this initial paragraph about the study system, as detailed below.

First of all, it is unclear whether this population shows breeding dispersal. The following information must be provided in order to make a compelling argument that breeding dispersal does not affect your conclusions: 1) what was the age at first breeding of all individuals included in the analyses of longevity and breeding success? 2) what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?

You should also made clear that you are focusing on a very specific sample of individuals, i.e. local recruits, which represent a tiny fraction of the total fledglings from this population. I am not saying this can introduce bias, but we know that these individuals may be somewhat odd, as the vast majority show natal dispersal, and there is evidence from other species that local recruits may differ from dispersing recruits in many different ways, especially in species where local recruitment is very low; they are extreme individuals, so it is unclear how patterns detected in these individuals reflect those occurring in the population at large. A word of caution should be added. Moreover, is natal dispersal/local recruitment sex biased? In many passerines, recruits are mostly males. What is the sex ratio of your local recruit population? Did you compare longevity of local recruits (males, females) with that of immigrant individuals (for which you should have data)?

Clearly, providing this information is essential to convince the reader that you have measured both longevity and LRT in a robust way.

Finally, you need to report in this section in detail how you have calculated fitness components (longevity and LRT). These variables are only first mentioned in the statistical analysis section, while they should be introduced and described here.

L175: I am not sure whether I would include plate ID in models as a fixed effect. Basically, it is a nuisance variable you aim at controlling for, you do not wish to test specific hypotheses about this variable. Why not including it as a random effect instead? An easier alternative would also be to remove plate effects by within-group centering (i.e. standardizing plate values to the reference sample value, mentioned in L154).

L182: I guess you meant 'a by-subject random slope with age' here, right? I am skeptical it is feasible to model a random slope age effect at the individual level when you have a maximum of 4 datapoints per subject (if I understood correctly). Please check this out. Nevertheless, this model shows that, despite the population shows a (non-significant!) tendency for a decrease in TL with age, individuals significantly differ in their age-related variation in TL, with some increasing and others decreasing. This individual variation and its implications for this study (if any) is not mentioned in the results.

L185: rephrase: ...freedom estimated according to the Kenward-Rogers method, and random effects...

L192: this is confusing. Of course if you had local recruits they must have bred at least once. Using the 'local recruit' definition would make everything clear. You should rather state here that birds were blood-sampled at least once upon settling as breeders in the population.

M L195 and subsequent. You should made it clear that you have verified a key assumption of Cox regression, i.e. a constant multiplicative effect in the hazard function over time for each covariate. Currently, this cannot be assessed (not even visually) as you are nowhere showing a survival curve. I urge you to present survival curves (e.g. in relation to sex) and to explicitly verify that Cox proportional hazard model assumptions are met. Otherwise, you can consider running a GLM with longevity as a dependent variable and sex and TL10 as predictors. This is the way you are showing it in Fig. 2A, by the way....of course in that case you should take care of the distribution of longevity (likely not purely Gaussian).

L198-205: move in first subheading of methods (see previous comment). Also, specify how you could estimate brood size (at which age? how much before fledging?). More details are needed here to properly assess the robustness of your results.

L211-212: the standard term for this trait is 'laying date'. Please refer to laying date throughout (instead of date of first egg, etc.).

L213: please refer to 'dataset', not 'datafile' (here and elsewhere).

L214-215 and subsequent: very poorly presented and confusing. What do you exactly mean by 'included some known covariates'? Please report clearly which covariates and which interactions were included in each model.

L216-217: year should be included as a random rather than fixed effect. You are not interesting in testing hypotheses about year differences.

L221-222: 'assuming' is not the best word here. You can test these 'assumptions' and check which distribution fits best your data. Sounds weird that Clutch size is Gaussian and fledging success is negbin. Please check it out. Then, once more, be more coherent with naming your variables, because it seems you are confusing fledging success with number of offspring. I agree that modelling the number of offspring with clutch size as an offset in practice equals to analyzing fledging success, but it is not correct to say that you analysed fledging success. Moreover, the standard term for 'number of offspring' is brood size. Please use this term throughout.

L231-232: actually, this is unclear to me because this formula does not refer to a 'bivariate' LMM, where TL and fitness are both dependent variables. What I see is that the two LMM (for TL and for LRT/longevity) here have different formulas, and I am unsure whether this is ok. Overall, I'm unfamiliar with BLMM, so I admit I may be wrong here. However, I am aware that the general recommendation is to avoid relying on BLMM unless you have a very large datasets with many repeated observations per individuals, which should be preferably collected at the same time point (not the case here, right?). I would thus avoid using this approach and remove these analyses from the manuscript (although they might be moved to the suppl. mat). Indeed, on the one hand you manage to avoid performing stats on stats, but on the other hand BLMM may lead to spurious conclusions as assumptions may be easily violated and power is generally low unless very large samples are available (see Dingenmanse and Dochtermann JAE). I would personally trust more results deriving from 'stats on stats' in this case...

L271: although significant, it should be specified that repeatability is definitely low.

L274: this statement is questionable given 1) the scatter shown in Fig. 1 and 2) the p-value of 0.053 for the age effect.

L275-276: this way of presenting results is confusing, in my opinion. Here you are testing the effect of sex, not F-M. Rather, specify in the methods that you have coded sex as 0 = female and 1 = male (correct?). What does F-M means? Difference in estimate between female and male values?

L278-280: be careful because the correlation of estimated slopes and intercepts may not have any actual biological value (https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#singular-models-random-effect-variances-estimated-as-zero-or-correlations-estimated-as---1) as it may depend on lack of convergence or other issues with model specifications. I recall reading some comments by mixed model 'gurus' stating that high correlation generally reflect model mis-specification and collinearity among predictors (which may considerably change by using e.g. centered predictors). I would personally avoid mentioning or interpreting these stats.

L283-284: this was puzzling to me. How could you estimate the change in TL with age for birds that did not breed? Were they captured during the non-breeding season when birds from nearby breeding sites gather together while moving to breeding sites which were not monitored? This requires some additional explanation in the Methods.

L292-293: again, a confusing way of presenting results from statistical analysis. Please be explicit in stating that you are testing an interaction term here (sex x TL10 on longevity).

L308: less than in females

L333-334: I find confusing here referring to 'lifetime success' and including longevity. I think you should always better separate LRS from longevity.

L333-335: where is the evidence for lack of sex differences? Please add reference to Table 3.

L339: again, a confusing mention of 'lifetime fitness' here. See previous comment. Be consistent with terminology throughout.

L411: wording unclear, rephrase: ....across species showing disparate life histories....groups remains to be elucidated.

L416: correct typo

Fig. 1: clarify what 'lines' are here. It seems these are 'regression lines' for each individual. I would rather show lines connecting individual datapoints for a given individual (it does not make much sense to compute slopes based on 4 datapoints...how could you assume the relationship is 'linear'?). BTW, this figure makes me wonder whether it would be worth using TL10 as a log-transformed variable, as it seems it is highly skewed. Also including it as a predictor may lead extreme values to have a high leverage on regression/mixed model output. I am unsure whether you did test it, but I would consider this carefully.

Fig. 2: confusing, as I do not think presenting separate 'linear regressions' is appropriate here. Moreover, I guess reference to GLMM here is wrong, as you have a single datapoint per individual, right? (see L 207-209). Please correct to GLM. And you should include confidence bands for each group (can be easily obtained via the 'visreg' R package). The best would be to plot 'raw' data with GLM-estimated slopes by group.

Table 1: statistics not properly reported. Under the random effects, Intercept should read 'Individual identity'. I also believe that 'plate' should be a further random intercept effect (or that it should be treated in a different way, as I suggested).

Table 2: confusing, requires improvement. Male-female should be Sex. Interactions are reported in a confusing way. Moreover, if year is included as a fixed effect (which I do not agree, by the way...why not as a random effect?), the statistics (global F test) should be reported. Why 'Subject age'? Never heard of this variable. Do you mean 'Age'? What do you mean by Telomere length at day 10 (Female)? And by Telomere at day 10 (M-F)? Also, list all predictors in the same order in all sub-tables, at present they are confusing. Why 'Age' is not included in models of clutch size and Offspring? To sum up, you need to present these results in an inconsistent and confusing manner which needs to be improved. Factor by covariate interactions should be reported as e.g. Sex ×TL10, with global F-test and associated estimates (one for each sex category), to highlight how slopes diverge. Check out the excellent 'reghelper' R package, command 'simple_slopes' to get group-specific slopes easily without too many calculations and recoding. The 'usual' way of reporting differences from a reference value is misleading and unclear.

Table 2: confusing, requires improvement. See previous comments about 'lifetime success' and how it generates confusion. The meaning of the variables (Females and Male-Female) are unclear.

Also, chi square tests are unclear here: what do they represent? I would expect to see one slope for males and one for females. Where do these statistics come from? Which kind of models? Which variables are included? These are not properly described in the methods. Are models ran separately for each sex? What is the sample size per sex? These details must be provided.

Referee: 2

Comments to the Author(s)
This is a well written, carefully thought-out, and comprehensively conducted analysis. The fact that the analysis script and data are included is a further laudable strength. I also appreciate the model robustness checks with supplemental Bayesian models (although I did not evaluate the Bayesian models). This should be published after minor revision. My specific suggestions below are all pretty minor points.

Title: if there is space would be good to fit in the species name

L53 – TLs shorten with age and predict longevity in SOME species—both of these patterns appear to be substantially heterogeneous across species

L62 – While this sentence is justified based on the broader literature—the findings of Epel et al have not held up to replication, and I would suggest citing a meta-analysis or seminal finding which does hold up. Same point applies to line 65—while a few studies do show reproduction predicting shorter TL in women, the Epel study is looking at caregiving for disabled children, which is not the same thing—and has not replicated well

L68 – TL loss suggests longitudinal change was measured—which was not done in all of these studies

Telomere length measurement methods are generally very well described and in much more detail than is typically provided. Nonetheless, please be sure all portions of the emerging best practices from the telomere research network are considered: https://osf.io/9pzst/

In particular, T/S ratios should be transformed to z-scores to make effects more interpretable following the TRN and Verhulst 2020 suggestions.

L134 - While GAPDH only having one amplicon size as discerned by melt curve and electrophoresis is a good check, it is still possible that the amplicon varies in copy number across individuals or has genetic polymorphisms which influence amplification efficiency. Such patterns would also tend to increase intra-individual repeatability of T/S ratios. This could be addressed by re-running samples with the highest and lowest GAPDH concentrations and trying another putative single-copy gene control to be sure it and the GAPDH results are similar. I don't think this experiment should be a requirement for publication, but would help to assure that TL is measured well here.

L152 – this is fine and reasonable. I would suggest that the authors consider using methods to adjust for efficiencies that deviate from the ideal of 2 (aka 100%) in their future work—but do not think this should be required for publication. While efficiencies were measured, they do not seem to have been used for calculating T/S

L158 – what were R^2 values of standard curve?

L161 – good—but please be more specific about which version of the ICC statistic was calculated here—there are several

Controlling for baseline TL can confound examination of longitudinal changes in TL (Bateson et al. 2019). Please double check that this error is not being made in your analyses.

L244 – I am a little concerned that ln transformation is changing the nature of the hypothesis being tested in a subtle way—that is you are now testing how TL predicts ln number of eggs and ln longevity which is a non-linear and biologically different hypothesis.

Figure 1 – nice and helpful figure—I suggest the authors consider making the lines and dots narrower/smaller, to allow readers to better discern the patterns

L278 - Greater variance in TL change in females is an interesting finding. Can you put a p value to this difference between sexes and/or show confidence intervals of these estimates so we can get a better sense for how different these are when considering the uncertainty of the estimates? Also, I don't think this finding was further considered in the discussion and probably should be.

L293 – would be easier to read if terminology was kept more consistent across the text and figures. It appears longevity, probability of disappearing and cumulative survival all have identical meanings—or if there is a distinction being made between these, it is not clear to me.

L329 – please make more clear in the text here the nature of the collinearity? i.e. does longer baseline TL predict greater decline in TL consistent with regression to the mean expectations?

It occurs to me that if it is blood TL which is having the effects on biology than blood TL measured later in life should be more predictive of reproductive success and longevity AFTER that TL measure than the earliest TL measure. I don't think this was directly tested in this paper. Given how comprehensive the analysis already is, I think the authors should feel free to test this or not at their own discretion.

Bateson M, Eisenberg DTA, and Nettle D. 2019. Controlling for baseline telomere length biases estimates of the rate of telomere attrition. Royal Society Open Science 6(10):190937.

Author's Response to Decision Letter for (RSPB-2020-3219.R0)

See Appendix A.

RSPB-2021-0560.R0

Review form: Reviewer 1

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
N/A

Is it clear?
N/A

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors have carefully revised their manuscript, which is now somewhat improved compared to the previous version.

Yet, some key questions remain open, in my opinion.

1) I disagree that using brood size instead of 'number of offspring' would make this manuscript less attractive or suitable for a broad audience. Any researcher of animal taxa would understand the term 'litter size' (mainly for viviparous taxa) or 'brood size' (mainly for oviparous altricial taxa). I think that using the proper terminology is important in biological studies, and you are submitting this manuscript to a biological journal.
2) I am still unconvinced that including plate identity in models as a fixed effect is suitable for controlling for among-plate differences in TL. Although I understand that random factors do not completely control for this, and that fixed effects may be more effective, the best alternative in my opinion would be mean-centering every value with respect to the plate reference. Mean-centering is not doing stat-on-stat (subtracting from a reference value does not imply performing any statistical analysis or estimate). I am still unconvinced that allowing for covariation between plate identity and the other variables in the model (as it is now) is appropriate, as it may mask real biological differences between plates (in terms of sample composition). Please note that I only wish to be sure that the results do not reflect statistical artifacts, which I think with the analysis of telomere variation is always a real possibility that must be ruled out to the extent to which this is possible.
3) You did not provide an adequate answer to one of my previous comments, neither in the response nor in the text, concerning the accuracy of LRS estimates for each individual in the sample (if I did not miss something). I had indeed asked "what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?". I could find no clear answer to this question in the response. My point is that, if the chances of missing one of the reproductive events in the lifetime of some individuals are high (for any reason), then you might not end up with an accurate estimate of LRS. Again, maybe I am missing something here and you do have breeding data for all individuals included in the sample and for all years

until disappearance (i.e. death), but then I would say that explicitly. But if this is not the case, and if one bird alive for 5 years has raised, i.e., 3, 4, -, 4, 5 offspring, but was not observed in year 3, this would seriously undermine the robustness of your conclusions. Of course this depends on the frequency of such missed reproductive events (if any). If this is a rare occurrence, it may not matter. But please provide an indication of the frequency of these events in your population. It is unclear what do you mean exactly by "Some breeding dispersal occurs; movement between locations is more common for females than males, but is still rare (<10%)". What matters here is the probability of missing a reproductive event in the lifetime of an individual, be it due to breeding dispersal or other events (e.g. monitoring flaws), and how does it affect your findings by potentially biasing estimates of LRS. At the very least, you should try to redo the analyses on the subset of individuals for which you have the complete breeding history (this information should also be provided). As an alternative, you can keep all birds in the analyses and assign weight 1 ('weight' statement in the R lme4 syntax) to observations with complete data, and proportionally smaller weight to observations with less complete reproductive histories (e.g. in the case above, the bird with 4/5 reproductive events should have a weight of 0.8). This would also correct for differences between the sexes in breeding dispersal.

4) The scatter of total offspring in Figure 2 makes me wonder about the scatter in the LRS data, which looks really huge. What is the effect of the single extreme individual on regression coefficients and model output? How are the results of the interaction terms affected by removing this single individual?

5) I think it would also be important to add to the results some basic descriptive statistics of your sample of individuals in terms of longevity and LRS, including sample size (this information could be added to Table 2 or 3), for both males and females. Also, be careful with the use of symbols, do not use X instead of the Greek letter for chi-square statistics (check throughout).

As a final minor remark, please avoid or limit the usage of the term 'critical*' throughout the ms,

Decision letter (RSPB-2021-0560.R0)

18-Apr-2021

Dear Dr Heidinger:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Editors have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" in the "File Upload" section. This should document, point by point, how you have responded to the

reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that

the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Locke Rowe
mailto: proceedingsb@royalsociety.org

Associate Editor Board Member
Comments to Author:
In this revised version of the manuscript, the authors have improved the clarity at several points, but as you can see in the comments from the reviewers, some questions still remain. While I am satisfied with the use of plate as a fixed effect, I agree that this version does not clearly address the prior concern about how the accuracy of lifetime reproductive success estimates might affect the results. I also agree that the addition of some descriptive statistics would be helpful.

In reading this version of the manuscript, I found myself wondering whether it might be more straightforward to use lifetime reproductive success throughout rather than lifetime fitness and lifetime reproductive success. While I do understand that there are components of fitness that can extend beyond reproductive success, in this study you principally focus on longevity and reproductive success. I would point you to lines 96-109, where the paragraph opens with "telomeres and lifetime fitness" and then midway through "two major hypotheses about links between telomere dynamics and lifetime reproductive success". If the hypotheses you are testing focus on reproductive success, simplifying the language would be useful.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s).
The authors have carefully revised their manuscript, which is now somewhat improved compared to the previous version.

Yet, some key questions remain open, in my opinion.

1) I disagree that using brood size instead of 'number of offspring' would make this manuscript less attractive or suitable for a broad audience. Any researcher of animal taxa would understand the term 'litter size' (mainly for viviparous taxa) or 'brood size' (mainly for oviparous altricial taxa). I think that using the proper terminology is important in biological studies, and you are submitting this manuscript to a biological journal.
2) I am still unconvinced that including plate identity in models as a fixed effect is suitable for controlling for among-plate differences in TL. Although I understand that random factors do not completely control for this, and that fixed effects may be more effective, the best alternative in my opinion would be mean-centering every value with respect to the plate reference. Mean-centering is not doing stat-on-stat (subtracting from a reference value does not imply performing any statistical analysis or estimate). I am still unconvinced that allowing for covariation between plate identity and the other variables in the model (as it is now) is appropriate, as it may mask real biological differences between plates (in terms of sample composition). Please note that I only wish to be sure that the results do not reflect statistical artifacts, which I think with the analysis of telomere variation is always a real possibility that must be ruled out to the extent to which this is possible.

3) You did not provide an adequate answer to one of my previous comments, neither in the response nor in the text, concerning the accuracy of LRS estimates for each individual in the sample (if I did not miss something). I had indeed asked “what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?”. I could find no clear answer to this question in the response. My point is that, if the chances of missing one of the reproductive events in the lifetime of some individuals are high (for any reason), then you might not end up with an accurate estimate of LRS. Again, maybe I am missing something here and you do have breeding data for all individuals included in the sample and for all years until disappearance (i.e. death), but then I would say that explicitly. But if this is not the case, and if one bird alive for 5 years has raised, i.e., 3, 4, -, 4, 5 offspring, but was not observed in year 3, this would seriously undermine the robustness of your conclusions. Of course this depends on the frequency of such missed reproductive events (if any). If this is a rare occurrence, it may not matter. But please provide an indication of the frequency of these events in your population. It is unclear what do you mean exactly by “Some breeding dispersal occurs; movement between locations is more common for females than males, but is still rare (<10%)”. What matters here is the probability of missing a reproductive event in the lifetime of an individual, be it due to breeding dispersal or other events (e.g. monitoring flaws), and how does it affect your findings by potentially biasing estimates of LRS. At the very least, you should try to redo the analyses on the subset of individuals for which you have the complete breeding history (this information should also be provided). As an alternative, you can keep all birds in the analyses and assign weight 1 (‘weight’ statement in the R lme4 syntax) to observations with complete data, and proportionally smaller weight to observations with less complete reproductive histories (e.g. in the case above, the bird with 4/5 reproductive events should have a weight of 0.8). This would also correct for differences between the sexes in breeding dispersal.
4) The scatter of total offspring in Figure 2 makes me wonder about the scatter in the LRS data, which looks really huge. What is the effect of the single extreme individual on regression coefficients and model output? How are the results of the interaction terms affected by removing this single individual?
5) I think it would also be important to add to the results some basic descriptive statistics of your sample of individuals in terms of longevity and LRS, including sample size (this information could be added to Table 2 or 3), for both males and females. Also, be careful with the use of symbols, do not use X instead of the Greek letter for chi-square statistics (check throughout).

As a final minor remark, please avoid or limit the usage of the term ‘critical*’ throughout the ms,

Author's Response to Decision Letter for (RSPB-2021-0560.R0)

See Appendix B.

Decision letter (RSPB-2021-0560.R1)

04-May-2021

Dear Dr Heidinger

I am pleased to inform you that your manuscript entitled "Longer telomeres during early life predict higher lifetime reproductive success in females but not males" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Data Accessibility section
Please remember to make any data sets live prior to publication, and update any links as needed when you receive a proof to check. It is good practice to also add data sets to your reference list.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Dr Locke Rowe
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
I thank the authors for their careful revision and have no additional changes.

Appendix A

Dear Editor Dr. Locke Rowe,

Thank you for this opportunity to submit a greatly modified version of our original manuscript entitled **"Longer telomeres during early life predict higher lifetime fitness in females but not males"** to be considered for publication in Proceedings of the Royal Society B. We appreciate the time and effort the associate editor and reviewers have put into our manuscript and their comments universally helped us improve the manuscript. We carefully considered all suggestions and provide point-by-point replies and the details of our modifications below.

Sincerely,

Britt Heidinger and David Westneat (on behalf of all authors)

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)
The authors report that among-individual TL variation of wild house sparrow during early life positively predicts lifetime reproductive success, and that this effect was sex-specific (it emerged for males but not for females) and entirely mediated by differences in longevity. I think the ms is generally well-written and clear. I do have however some major concerns (labeled M), as I will detail below, that I urge the authors to carefully consider in order to make their case stronger. All comments are written in order of appearance in the ms.

L1: I found the title rather misleading upon carefully reading the ms. Actually, if longer telomeres non-significantly predict reproductive success upon controlling for longevity, stating 'predicts higher lifetime fitness' can lead to the false impression that TL predicts reproductive success. I would thus rephrase it as follows: "....predicts greater female but not male longevity"

We appreciate this perspective, but respectfully disagree. Numerous studies have demonstrated that individuals with longer telomeres live longer. But importantly, individuals that live longer, may or may not have higher lifetime fitness and this will depend on the number of offspring produced throughout life, which we aimed to test here. This study is unique in that we were able to follow individuals across their lifespans and measure not only longevity, but also lifetime reproductive success and various metrics of reproductive rates or performance. For this reason, we think the title more accurately reflects both what was measured and why this study makes a novel contribution. We make it clear, beginning in the abstract and throughout the manuscript that this positive relationship between telomeres and lifetime fitness is due to the relationship between telomere length and longevity and not reproductive output. We could potentially modify the title as follows "Longer telomeres during early life predict higher lifetime fitness in females but not males due to positive effects on longevity and not reproductive rate". However, this greatly increases the length of the title and there is evidence on the instructions for authors page of this journal that manuscripts with longer titles have a lower citation rate (https://royalsocietypublishing.org/doi/full/10.1098/rsos.150266) and we would thus prefer not to make this adjustment.

L33: the reference to ‘reproductive rates’ is somewhat obscure here and can be understood only upon reading carefully the results. I suggest rephrasing this sentence to make it clearer.

Thank you for this suggestion, this has now been modified by substituting “reproduction per year or attempt” on Line 34.

L75: telomeres being a correlated

Thank you for this suggestion, this has now been modified on line 76.

L76-77: suffer lower reproductive costs

Thank you for this suggestion, this has now been modified on line 78.

L81: few non-human studies have

Thank you for this suggestion, this has now been modified on line 82.

M L107 and subsequent: far more details are needed here about the study population characteristics and how you have calculated the fitness endpoints in this initial paragraph about the study system, as detailed below.

First of all, it is unclear whether this population shows breeding dispersal. The following information must be provided in order to make a compelling argument that breeding dispersal does not affect your conclusions: 1) what was the age at first breeding of all individuals included in the analyses of longevity and breeding success? 2) what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?

You should also made clear that you are focusing on a very specific sample of individuals, i.e. local recruits, which represent a tiny fraction of the total fledglings from this population. I am not saying this can introduce bias, but we know that these individuals may be somewhat odd, as the vast majority show natal dispersal, and there is evidence from other species that local recruits may differ from dispersing recruits in many different ways, especially in species where local recruitment is very low; they are extreme individuals, so it is unclear how patterns detected in these individuals reflect those occurring in the population at large. A word of caution should be added. Moreover, is natal dispersal/local recruitment sex biased? In many passerines, recruits are mostly males. What is the sex ratio of your local recruit population? Did you compare longevity of local recruits (males, females) with that of immigrant individuals (for which you should have data)?

Thank you for these suggestions. We have added new information in several places:

Lines 116-117: Information about recruitment of both sexes (the sex-ratio at 10 days is nearly 50:50, Westneat et al. 2003). It is not appropriate to supply vague unpublished data in the main text, but the numbers are 204 females returned to breed and 225 males did so. In addition, we added a line (119-121) that most recruits breed in their first adult year (81% of females, 58% of males, lower because “breeding” requires eggs being laid but some males advertised without pairing with a female).

Clearly, providing this information is essential to convince the reader that you have measured both longevity and LRT in a robust way.

Finally, you need to report in this section in detail how you have calculated fitness components (longevity and LRT). These variables are only first mentioned in the statistical analysis section, while they should be introduced and described here.

Line 121-125: We added information about the measures of lifetime fitness components.

L175: I am not sure whether I would include plate ID in models as a fixed effect. Basically, it is a nuisance variable you aim at controlling for, you do not wish to test specific hypotheses about this variable. Why not including it as a random effect instead? An easier alternative would also be to remove plate effects by within-group centering (i.e. standardizing plate values to the reference sample value, mentioned in L154).

Although we agree that plate is a nuisance variable that is irrelevant to any hypothesis we are testing, we do not think it is more appropriate to treat this variable as a random rather than a fixed effect. This is because random effects partition variance and so remove it from the residual or other random effects, but they do not control for it. Any fixed effect, such as sex, which likely varies within and among plates to some degree, will end up being influenced by this spurious variance if it is merely confined to a random effect. Indeed, the problem of disentangling variables that have effects at two or more levels is a major challenge (van de Pol and Wright 2009, Westneat et al. 2020) that is not solved by putting factors into the random effects. By contrast, using plate as a fixed effect adjusts all the data to the mean of the referent plate. This is preferable because all other fixed effects and the other random effects are then assessed using the residuals from mean plate values. This is also much better than adjusting such values before analysis, as it avoids stats-on-stats problems of the sort noted by Freckleton (Freckleton, Robert P. "On the misuse of residuals in ecology: regression of residuals vs. multiple regression." Journal of Animal Ecology 71.3 (2002): 542-545.) because any uncertainty in the referent is never accounted for in that approach.

L182: I guess you meant ‘a by-subject random slope with age’ here, right? I am skeptical it is feasible to model a random slope age effect at the individual level when you have a maximum of 4 datapoints per subject (if I understood correctly). Please check this out. Nevertheless, this model shows that, despite the population shows a (non-significant!) tendency for a decrease in TL with age, individuals significantly differ in their age-related variation in TL, with some increasing and others decreasing. This individual variation and its implications for this study (if any) is not mentioned in the results.

We changed the wording of this on line 190. Although a sample of 4 data points per individual (max) reduces the power of a test of random slopes, it does not make the analysis impossible. Only 3 data points are required to get a slope and have residual variation in order to test it. We do not make too much of the random slope variation because slope is collinear with the intercept, thus making them impossible to really separate. This is probably driven by the small sample of individuals with 3 or more measures and we present this result in three places: line 283, in Table 1 (“Slope:Age”), and Figure 1.

L185: rephrase: ...freedom estimated according to the Kenward-Rogers method, and random effects...

Thank you for this suggestion, this has now been modified on line 193.

L192: this is confusing. Of course if you had local recruits they must have bred at least once. Using the

‘local recruit’ definition would make everything clear. You should rather state here that birds were blood-sampled at least once upon settling as breeders in the population.

Thank you, this has now been modified on line 199-200.

M L195 and subsequent. You should made it clear that you have verified a key assumption of Cox regression, i.e. a constant multiplicative effect in the hazard function over time for each covariate. Currently, this cannot be assessed (not even visually) as you are nowhere showing a survival curve. I urge you to present survival curves (e.g. in relation to sex) and to explicitly verify that Cox proportional hazard model assumptions are met. Otherwise, you can consider running a GLM with longevity as a dependent variable and sex and TL10 as predictors. This is the way you are showing it in Fig. 2A, by the way....of course in that case you should take care of the distribution of longevity (likely not purely Gaussian).

Thank you for this suggestion. We have presented the survival curves by sex in the supplemental material, and we now also include tests of the proportionality assumption. There was no evidence that any of the variables deviated from that assumption; this is now referred to in the main text (lines 206-207) and explained in more detail in the supplement.

L198-205: move in first subheading of methods (see previous comment). Also, specify how you could estimate brood size (at which age? how much before fledging?). More details are needed here to properly assess the robustness of your results.

Thank you for this suggestion. As noted above, we added details earlier.

L211-212: the standard term for this trait is ‘laying date’. Please refer to laying date throughout (instead of date of first egg, etc.).

Thank you for calling this to our attention. In house sparrows, the term laying date is misleading because birds can produce multiple clutches across the season (i.e., laying eggs up to 6 different times in a season and therefore having 6 laying dates in a year). We reviewed our use of this term throughout the manuscript to better distinguish the lay date for each nest versus the date of first breeding in the season.

L213: please refer to ‘dataset’, not ‘datafile’ (here and elsewhere).

Thank you for this suggestion, this has now been modified throughout the manuscript.

L214-215 and subsequent: very poorly presented and confusing. What do you exactly mean by ‘included some known covariates’? Please report clearly which covariates and which interactions were included in each model.

Thank you for this suggestion. We have now altered the wording on lines 224-231.

L216-217: year should be included as a random rather than fixed effect. You are not interesting in testing hypotheses about year differences.

As described above regarding plate, we disagree that year should be included as a random rather than a fixed effect in our analyses. If year is treated as a random effect, this pulls variation out of other variance components, but any fixed effect that varies within and across years would nevertheless be included in that variance. As with plate above, our intent here is to factor out the variance due to year in order to assess success for birds who lived across years in a study that spanned many more years than they live. By employing year as a fixed effect, all other fixed effects are tested using the residuals from year. This

does mean that each individual event within a year is treated independently, but because the main hypothesis is about how telomere length at Day 10 influences breeding performance, and the day 10 measure is in a different year than the breeding data, and because year of hatch has no effect on telomere length, there is no reason to think that there is pseudoreplication within years.

L221-222: 'assuming' is not the best word here. You can test these 'assumptions' and check which distribution fits best your data. Sounds weird that Clutch size is Gaussian and fledging success is negbin. Please check it out.

We altered the "assuming" wording (now lines 229-230), but note that even with tests, one still assumes the underlying distribution (non-significance does not mean that is the actual distribution). What distribution is appropriate deserves some explanation. Clutch size fits no known distribution but is closest to Gaussian (Westneat et al. 2009, 2014). The reason is that each egg in sequence is laid with very different probability (in sparrows, it is 100% by definition for the first egg, and nearly 100% for eggs 2-3, then drops a bit for egg 4, more for 5, and then plummets after that, reaching 0 at about 8 eggs. So, Poisson is definitely inappropriate and negative binomial does not fit either. We note that with large sample sizes, nearly all data will deviate from the theoretical distribution types, but luckily such deviations do not matter much at all (Schielzeth et al. 2020). Offspring number is modeled as a negative binomial because of many more probabilistic events that lead to the final count and due to major dispersion in the distribution which makes it unsuitable for Poisson.

Then, once more, be more coherent with naming your variables, because it seems you are confusing fledging success with number of offspring. I agree that modelling the number of offspring with clutch size as an offset in practice equals to analyzing fledging success, but it is not correct to say that you analysed fledging success. Moreover, the standard term for 'number of offspring' is brood size. Please use this term throughout.

Thank you for this suggestion, we have now altered to wording to make this clearer. We resisted using "brood size" as it is a taxonomically restricted term (birds, fish, and social insects) and prefer "number of offspring" to retain accessibility to those studying other organisms.

L231-232: actually, this is unclear to me because this formula does not refer to a 'bivariate' LMM, where TL and fitness are both dependent variables. What I see is that the two LMM (for TL and for LRT/longevity) here have different formulas, and I am unsure whether this is ok. Overall, I'm unfamiliar with BLMM, so I admit I may be wrong here. However, I am aware that the general recommendation is to avoid relying on BLMM unless you have a very large datasets with many repeated observations per individuals, which should be preferably collected at the same time point (not the case here, right?). I would thus avoid using this approach and remove these analyses from the manuscript (although they might be moved to the suppl. mat). Indeed, on the one hand you manage to avoid performing stats on stats, but on the other hand BLMM may lead to spurious conclusions as assumptions may be easily violated and power is generally low unless very large samples are available (see Dingenmanse and Dochtermann JAE). I would personally trust more results deriving from 'stats on stats' in this case...

You have raised several important issues here. First, we reformatted the equations in the text to make it clearer that this is a bivariate model (lines 241-246). Second, separate equations are perfectly fine for doing multivariate models, either using Proc Mixed in SAS or brms in r. In fact, brms is explicitly designed for separate equations and, like SAS, can accommodate different distributional assumptions as well. Multivariate models are data hungry for some parameters but not others, so it depends on which ones one wants to assess. As we point out, some of the potential benefits of doing these models could not be achieved because of limited data (we can't assess covariance between intercept or slope and longevity independently because of the inability to distinguish intercept and slope). Nevertheless, the widespread

advantage of this method is that it avoids all the problems with various types of stats on stats, noted by the reviewer. Finally, Dingemanse & Dochtermann's supplement (S17.C) illustrates ways of assessing traits that are not measured at the same time. Clearly, the residual covariances cannot be calculated, but the G-side covariances (in this case, among-individuals) can be and this is the focus of our analysis.

L271: although significant, it should be specified that repeatability is definitely low.

We are amenable to modifying the text to suggest that the repeatability is low, but are uncertain why this suggestion is being made or what our value is being compared to. We note that many estimates of repeatability are either done incorrectly (by including inappropriate fixed effects), or are assessing repeatability over very different time frames than ours. We are not sure a complex comparison in our results section is warranted.

L274: this statement is questionable given 1) the scatter shown in Fig. 1 and 2) the p-value of 0.053 for the age effect.

Yet the statistical tests show the random slope addition to the model is significant and the idea of them is important so we have left this as is.

L275-276: this way of presenting results is confusing, in my opinion. Here you are testing the effect of sex, not F-M. Rather, specify in the methods that you have coded sex as 0 = female and 1 = male (correct?).

Sex is coded as "M" and "F" and is a factor variable. In this case, the analyses treated Female as the referent sex.

What does F-M means? Difference in estimate between female and male values?

Yes, the effect size of sex will always be the difference between them. Without this nomenclature, the effect size is uninterpretable. We altered the wording to clarify (lines 284-286).

L278-280: be careful because the correlation of estimated slopes and intercepts may not have any actual biological value (https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#singular-models-random-effect-variances-estimated-as-zero-or-correlations-estimated-as---1) as it may depend on lack of convergence or other issues with model specifications. I recall reading some comments by mixed model 'gurus' stating that high correlation generally reflect model mis-specification and collinearity among predictors (which may considerably change by using e.g. centered predictors). I would personally avoid mentioning or interpreting these stats.

We agree that this needs to be interpreted cautiously and we made every effort to do so. The reason for the high covariance is not due to collinearity of predictors, since there is only 1 in the model (age). We suspect this is a consequence of relatively few individuals with 3 or more data points and probably reflects the impact of regression to the mean more than anything else. But, this is an outcome that could have been different and is therefore useful to present both because it is conceptually important and the result is useful to have, but we do not make very much of it through the rest of the paper.

L283-284: this was puzzling to me. How could you estimate the change in TL with age for birds that did not breed? Were they captured during the non-breeding season when birds from nearby breeding sites gather together while moving to breeding sites which were not monitored? This requires some additional explanation in the Methods.

We have now clarified in the methods (line 114) that birds were caught year round.

L292-293: again, a confusing way of presenting results from statistical analysis. Please be explicit in stating that you are testing an interaction term here (sex x TL10 on longevity).

Thank you for this suggestion, we have now modified this wording to improve clarity (line 312, 314, 316, and 319) but retaining the M-F to indicate the test of the difference between them.

L308: less than in females

Thank you for this suggestion, this has now been modified.

L333-334: I find confusing here referring to 'lifetime success' and including longevity. I think you should always better separate LRS from longevity.

We respectfully disagree. By definition, lifetime reproductive success must include longevity. This is why we distinguish lifetime reproductive success, longevity, and reproductive rate (reproduction per unit of time, therefore controlling for longevity) and reproductive performance (reproduction per attempt).

L333-335: where is the evidence for lack of sex differences? Please add reference to Table 3.

Thank you for this suggestion, this has now been modified (line 319).

L339: again, a confusing mention of 'lifetime fitness' here. See previous comment. Be consistent with terminology throughout.

Please see our previous responses to this issue above.

L411: wording unclear, rephrase: ....across species showing disparate life histories....groups remains to be elucidated.

Thank you for this suggestion, this has now been modified on lines 421-423.

L416: correct typo

Thank you for this suggestion, this has now been modified (lines 426-427).

Fig. 1: clarify what 'lines' are here. It seems these are 'regression lines' for each individual. I would rather show lines connecting individual datapoints for a given individual (it does not make much sense to compute slopes based on 4 datapoints...how could you assume the relationship is 'linear'?).

Thank you, we have clarified the legend to indicate these are linear regressions. Connecting dots directly would be a mess, so we have left this as is given that all the analyses in the paper assess linearity as well.

BTW, this figure makes me wonder whether it would be worth using TL10 as a log-transformed variable, as it seems it is highly skewed. Also including it as a predictor may lead extreme values to have a high leverage on regression/mixed model output. I am unsure whether you did test it, but I would consider this carefully.

Thank you for the observation. There is indeed some skew here, but nothing approach the skew of longevity. ln transformation changes the nature of the biological process being tested (see reviewer 2's

comment below and our response there) and makes interpreting effect sizes difficult. The skew that is present in T/S ratios is well within bounds of having little effect on estimates in mixed models (Schielzeth et al. 2020). There was no evidence of any undue leverage.

Fig. 2: confusing, as I do not think presenting separate 'linear regressions' is appropriate here. Moreover, I guess reference to GLMM here is wrong, as you have a single datapoint per individual, right? (see L 207-209). Please correct to GLM. And you should include confidence bands for each group (can be easily obtained via the 'visreg' R package). The best would be to plot 'raw' data with GLM-estimated slopes by group.

Thank you for these suggestions. We have modified the figure to include confidence bands. The statistical model testing these slopes only included plate besides age, so there is not much difference between the separate regressions here and those obtained from the full model. We also feel strongly that displaying the raw data is important. Finally, after messing around with visreg some, we decided that ggplot allowed us to much more finely control the appearance of the plot. We hope the revised figure is satisfactory.

Table 1: statistics not properly reported. Under the random effects, Intercept should read 'Individual identity'. I also believe that 'plate' should be a further random intercept effect (or that it should be treated in a different way, as I suggested).

Thank you for this suggestion. We have altered the labels to be clearer, but an issue is that Individual identity is present twice, since there is also an individual slope term. We left the term intercept in to make this distinction clearer.

Table 2: confusing, requires improvement. Male-female should be Sex. Interactions are reported in a confusing way. Moreover, if year is included as a fixed effect (which I do not agree, by the way...why not as a random effect?), the statistics (global F test) should be reported. Why 'Subject age'? Never heard of this variable. Do you mean 'Age'? What do you mean by Telomere length at day 10 (Female)? And by Telomere at day 10 (M-F)? Also, list all predictors in the same order in all sub-tables, at present they are confusing. Why 'Age' is not included in models of clutch size and Offspring? To sum up, you need to present these results in an inconsistent and confusing manner which needs to be improved. Factor by covariate interactions should be reported as e.g. Sex ×TL10, with global F-test and associated estimates (one for each sex category), to highlight how slopes diverge. Check out the excellent 'reghelper' R package, command 'simple_slopes' to get group-specific slopes easily without too many calculations and recoding. The 'usual' way of reporting differences from a reference value is misleading and unclear.

Thank you for these suggestions. We have revised the table presentation extensively.

Table 3: confusing, requires improvement. See previous comments about 'lifetime success' and how it generates confusion. The meaning of the variables (Females and Male-Female) are unclear. Also, chi square tests are unclear here: what do they represent? I would expect to see one slope for males and one for females. Where do these statistics come from? Which kind of models? Which variables are included? These are not properly described in the methods. Are models ran separately for each sex? What is the sample size per sex? These details must be provided.

Thank you for this suggestion. We have revised the Table to provide more details and improve presentation.

Referee: 2

Comments to the Author(s)
This is a well written, carefully thought-out, and comprehensively conducted analysis. The fact that the analysis script and data are included is a further laudable strength. I also appreciate the model robustness checks with supplemental Bayesian models (although I did not evaluate the Bayesian models). This should be published after minor revision. My specific suggestions below are all pretty minor points.

Title: if there is space would be good to fit in the species name

We would prefer not to include the species name as it would increase the length of the title and there is evidence in the instructions for authors for this journal that manuscripts with longer titles get cited at a lower rate (https://royalsocietypublishing.org/doi/full/10.1098/rsos.150266).

L53 – TLs shorten with age and predict longevity in SOME species—both of these patterns appear to be substantially heterogeneous across species

We have now modified the text to read, "In support of these ideas, telomeres shorten with age in diverse tissues and their length often predicts subsequent longevity, although this pattern is not universal". Please see lines 52-54.

L62 – While this sentence is justified based on the broader literature—the findings of Epel et al have not held up to replication, and I would suggest citing a meta-analysis or seminal finding which does hold up. Same point applies to line 65—while a few studies do show reproduction predicting shorter TL in women, the Epel study is looking at caregiving for disabled children, which is not the same thing—and has not replicated well

Thank you for this suggestion. We have now changed the citation on line 62 to Chatelain et al. 2019 and the citation that on line 65 to Sudyka 2019.

L68 – TL loss suggests longitudinal change was measured—which was not done in all of these studies

Thank you for this suggestion, we have now modified this to "associated with greater telomere loss and/or shorter telomeres" on line 68.

Telomere length measurement methods are generally very well described and in much more detail than is typically provided. Nonetheless, please be sure all portions of the emerging best practices from the telomere research network are considered: https://osf.io/9pzst/

We appreciate the importance of this and have endeavored to report our methods according to these recommendations.

In particular, T/S ratios should be transformed to z-scores to make effects more interpretable following the TRN and Verhulst 2020 suggestions.

While we appreciate this suggestion, all of these data will be made publicly available, ensuring that future comparisons can be made across studies.

L134 - While GAPDH only having one amplicon size as discerned by melt curve and electrophoresis is a good check, it is still possible that the amplicon varies in copy number across individuals or has genetic polymorphisms which influence amplification efficiency. Such patterns would also tend to increase intra-individual repeatability of T/S ratios. This could be addressed by re-running samples with the highest and

lowest GAPDH concentrations and trying another putative single-copy gene control to be sure it and the GAPDH results are similar. I don't think this experiment should be a requirement for publication, but would help to assure that TL is measured well here.

Although we appreciate this theoretical concern, we have no reason to suspect that it is a practical issue here. We would have expected variation in copy number or genetic polymorphisms to have affected the melt curve analysis but saw no indication of this. It seems much more likely that any slight variations in amplification efficiency would be due to small variations in pipetting, which is exactly what GAPDH is meant to control for. In addition, we have a collaborator, who is sequencing genomes of many house sparrows across populations and has confirmed that the individuals he assayed have a single copy of GAPDH. For all of these reasons, we think it is highly unlikely that variation in copy number or polymorphisms of GAPDH have contributed to the patterns we report here.

L152 – this is fine and reasonable. I would suggest that the authors consider using methods to adjust for efficiencies that deviate from the ideal of 2 (aka 100%) in their future work—but do not think this should be required for publication. While efficiencies were measured, they do not seem to have been used for calculating T/S

Thank you for this suggestion, you are correct that the efficiencies were not used to adjust the T/S ratios here, but we will consider doing this in future studies.

L158 – what were R^2 values of standard curve?

The average R^2 values of the standard curves were: GAPDH: 0.987 ± SEM 0.003 and telomeres: 0.983 ± SEM 0.002.

L161 – good—but please be more specific about which version of the ICC statistic was calculated here—there are several

Thank you for this suggestion, we have added this information to the text on line 169-170.

Controlling for baseline TL can confound examination of longitudinal changes in TL (Bateson et al. 2019). Please double check that this error is not being made in your analyses.

We confirm this error was not made here.

L244 – I am a little concerned that ln transformation is changing the nature of the hypothesis being tested in a subtle way—that is you are now testing how TL predicts ln number of eggs and ln longevity which is a non-linear and biologically different hypothesis.

We appreciate your thoughts on this. We use lnLongevity throughout, even though Figure 2 and some effect sizes are given in the natural scale (months), so there is no change in how we have treated longevity. By contrast, we do not ln transform T/S ratio, in large part because the skew is much smaller. Longevity is greatly skewed and would create undue leverage of a few long-lived individuals on the analysis if it was done on the natural scale. While divining whether the relationship is linear vs non-linear would be interesting, we probably do not have the resolution to do this. Lifetime eggs and offspring are also ln transformed because of the magnitude of the skew.

Figure 1 – nice and helpful figure—I suggest the authors consider making the lines and dots narrower/smaller, to allow readers to better discern the patterns

Thank you for this suggestion, we have attempted to modify this accordingly.

L278 - Greater variance in TL change in females is an interesting finding. Can you put a p value to this difference between sexes and/or show confidence intervals of these estimates so we can get a better sense for how different these are when considering the uncertainty of the estimates? Also, I don't think this finding was further considered in the discussion and probably should be.

We have addressed this in the supplement, but opted not to discuss it further because it was not significant.

L293 – would be easier to read if terminology was kept more consistent across the text and figures. It appears longevity, probability of disappearing and cumulative survival all have identical meanings—or if there is a distinction being made between these, it is not clear to me.

Thank you for this suggestion, this has now been modified.

L329 – please make more clear in the text here the nature of the collinearity? i.e. does longer baseline TL predict greater decline in TL consistent with regression to the mean expectations?

Thank you for calling this to our attention. Although this information is presented on line 288-289, we have now modified the wording here to remind the readers (line 341).

It occurs to me that if it is blood TL which is having the effects on biology than blood TL measured later in life should be more predictive of reproductive success and longevity AFTER that TL measure than the earliest TL measure. I don't think this was directly tested in this paper. Given how comprehensive the analysis already is, I think the authors should feel free to test this or not at their own discretion.

Thank you for this interesting idea, but there are several reasons why we are unable to assess it here. Most importantly, the data we collected are not very amenable to this analysis since the samples from later in life were collected opportunistically and vary considerably in how temporally linked they are with the breeding efforts. In addition, telomere length is somewhat repeatable, so the telomere length of later samples are correlated with the telomere lengths of the 10 day samples we analyzed. Thus, it would be difficult with the data we have to assess if later samples have an independent effect on breeding or survival. Lastly, adding this idea to the ones we already pursued here would greatly increase the length of the manuscript.

Bateson M, Eisenberg DTA, and Nettle D. 2019. Controlling for baseline telomere length biases estimates of the rate of telomere attrition. Royal Society Open Science 6(10):190937.

Appendix B

Dear Editor Dr. Locke Rowe,

Thank you for this opportunity to submit a modified version of our original manuscript entitled **"Longer telomeres during early life predict higher lifetime reproductive success in females but not males"** to be considered for publication in *Proceedings of the Royal Society B*. We appreciate the time and effort the associate editor and reviewer have put into our manuscript. We carefully considered all suggestions, modified the manuscript accordingly, and provide point-by-point replies and the details of our modifications below.

Sincerely,

Britt Heidinger and David Westneat (on behalf of all authors)

Reviewer(s)' Comments to Author:

Associate editor:

In this revised version of the manuscript, the authors have improved the clarity at several points, but as you can see in the comments from the reviewers, some questions still remain. While I am satisfied with the use of plate as a fixed effect, I agree that this version does not clearly address the prior concern about how the accuracy of lifetime reproductive success estimates might affect the results. I also agree that the addition of some descriptive statistics would be helpful.

In reading this version of the manuscript, I found myself wondering whether it might be more straightforward to use lifetime reproductive success throughout rather than lifetime fitness and lifetime reproductive success. While I do understand that there are components of fitness that can extend beyond reproductive success, in this study you principally focus on longevity and reproductive success. I would point you to lines 96-109, where the paragraph opens with "telomeres and lifetime fitness" and then midway through "two major hypotheses about links between telomere dynamics and lifetime reproductive success". If the hypotheses you are testing focus on reproductive success, simplifying the language would be useful.

Thank you for this valuable feedback. We have now changed lifetime fitness to lifetime reproductive success in most places throughout (including the title) to try and improve clarity. We have also provided more information about the calculation and accuracy of our lifetime reproductive success measures and additional descriptive statistics (please see below in the response to the reviewer).

Referee: 1

1) I disagree that using brood size instead of 'number of offspring' would make this manuscript less attractive or suitable for a broad audience. Any researcher of animal taxa would understand the term 'litter size' (mainly for viviparous taxa) or 'brood size' (mainly for oviparous altricial taxa). I think that using the proper terminology is important in biological studies, and you are submitting this manuscript to a biological journal.

We have changed wording to use "brood size".

2) I am still unconvinced that including plate identity in models as a fixed effect is suitable for controlling

for among-plate differences in TL. Although I understand that random factors do not completely control for this, and that fixed effects may be more effective, the best alternative in my opinion would be mean-centering every value with respect to the plate reference. Mean-centering is not doing stat-on-stat (subtracting from a reference value does not imply performing any statistical analysis or estimate). I am still unconvinced that allowing for covariation between plate identity and the other variables in the model (as it is now) is appropriate, as it may mask real biological differences between plates (in terms of sample composition). Please note that I only wish to be sure that the results do not reflect statistical artifacts, which I think with the analysis of telomere variation is always a real possibility that must be ruled out to the extent to which this is possible.

We appreciate the reviewer's concern, but fortuitously, including plate as a fixed effect factor does mean-center all the data to the referent plate, and also additionally centers the data within a plate to the mean of the plate relative to the referent plate, thereby making all the data relative to the referent plate. It does this without assuming (as mean-centering before analysis would) that the mean of the referent plate is estimated without error. It is this estimation without error that makes using mean-centering before analysis an example of stats-on-stats (Freckleton, Robert P. "On the misuse of residuals in ecology: regression of residuals vs. multiple regression." Journal of Animal Ecology 71.3 (2002): 542-545.).

3) You did not provide an adequate answer to one of my previous comments, neither in the response nor in the text, concerning the accuracy of LRS estimates for each individual in the sample (if I did not miss something). I had indeed asked "what is the likelihood of missing reproductive events of individuals upon they have first bred in your study population?". I could find no clear answer to this question in the response. My point is that, if the chances of missing one of the reproductive events in the lifetime of some individuals are high (for any reason), then you might not end up with an accurate estimate of LRS. Again, maybe I am missing something here and you do have breeding data for all individuals included in the sample and for all years until disappearance (i.e. death), but then I would say that explicitly. But if this is not the case, and if one bird alive for 5 years has raised, i.e., 3, 4, -, 4, 5 offspring, but was not observed in year 3, this would seriously undermine the robustness of your conclusions. Of course this depends on the frequency of such missed reproductive events (if any). If this is a rare occurrence, it may not matter. But please provide an indication of the frequency of these events in your population. It is unclear what do you mean exactly by "Some breeding dispersal occurs; movement between locations is more common for females than males, but is still rare (<10%)". What matters here is the probability of missing a reproductive event in the lifetime of an individual, be it due to breeding dispersal or other events (e.g. monitoring flaws), and how does it affect your findings by potentially biasing estimates of LRS. At the very least, you should try to redo the analyses on the subset of individuals for which you have the complete breeding history (this information should also be provided). As an alternative, you can keep all birds in the analyses and assign weight 1 ('weight' statement in the R lme4 syntax) to observations with complete data, and proportionally smaller weight to observations with less complete reproductive histories (e.g. in the case above, the bird with 4/5 reproductive events should have a weight of 0.8). This would also correct for differences between the sexes in breeding dispersal.

We thank the reviewer for clarifying and agree this is a legitimate concern, although for only one (an important one) of our analyses. We show that our data are good for measuring longevity and reproductive performance, which means the main result that telomeres predict longevity for females but not males is robust. We tracked all (100%) of the breeding attempts in the focal boxes that produced complete clutches in all years. Once a bird has started breeding in our boxes, missing reproduction is very rare; only 2 (3.6%) of the males in our study and 3 (6.2%) of females have a gap in reproduction. If we weight these data as suggested and rerun the key analyses, the effects are even stronger. We left the results as they were but added more information into the methods.

4) The scatter of total offspring in Figure 2 makes me wonder about the scatter in the LRS data, which looks really huge. What is the effect of the single extreme individual on regression coefficients and model output? How are the results of the interaction terms affected by removing this single individual?

Although a judgement based on the standard error might not apply to non-Gaussian analysis (as was used here) this individual does have a leverage (0.29) in our analysis justifying taking a closer look. There is, however, no other reason to remove her from the analysis—since nearly all subjects in the analysis had complete information (as did this bird) and she did not have an abnormally high clutch size or fledging success—she just lived a long time—she is a viable data point. Her high leverage alters some results but not the main difference between the sexes in the relationship between telomeres and LRS, so we have retained her in the data.

5) I think it would also be important to add to the results some basic descriptive statistics of your sample of individuals in terms of longevity and LRS, including sample size (this information could be added to Table 2 or 3), for both males and females. Also, be careful with the use of symbols, do not use X instead of the Greek letter for chi-square statistics (check throughout).

The sample sizes were present in the text in several places and in the figure legends; we have now added them to the tables as well. We have checked the symbols; they were all capital chi, which may get altered when uploading.

As a final minor remark, please avoid or limit the usage of the term ‘critical*’ throughout the ms,

We have reduced the use of the word “critical”.