Dear Recommender,
Dear Reviewer,

We would like to thank you for the opportunity to revise and resubmit our Stage 1 registered report (RR), entitled “Neophobia across social contexts in juvenile Herring gulls”.

We are grateful for the insightful comments provided by you and the reviewers. We have addressed recurring feedback from multiple reviewers in a main response below and have added detailed, point-by-point responses to individual comments. These adjustments have also been incorporated into the revised manuscript.

We have incorporated almost all the comments and suggestions from the reviewers and the recommender in the revised manuscript. We have changed our experimental design and will assign nest mates to different groups for testing. We have also revised the duration of the task, following the suggestion that the 10-minute period should start as soon as the last bird enters the test arena, giving each bird 10 minutes to eat. We will also give them 10 minutes in the start area. Finally, we will perform only one analysis (instead of two), and we have also adjusted our power analysis to account for a potential 10% drop-out rate, as our rearing protocol does not allow for species identification prior to testing. It is clear from this analysis that our study retains sufficient statistical power to detect small effect sizes even after taking into account the exclusion of lesser black-backed gulls. Thank you also for pointing out textual errors in the manuscript and references. We have implemented all these suggestions as well.

We have also received three comments which we would like to respond to here in detail:

(i) Our initial prediction of reduced repeatability was indeed closely linked to the expectation of reduced variance. Our reasoning was that a decrease in variance within a group setting might not generalise across different contexts, leading to reduced repeatability when comparing individual to group settings. However, on reflection, and taking into account your feedback, we realise that this relationship may not have been adequately justified and could lead to confusion. We have therefore decided to drop the prediction of reduced repeatability across contexts. Instead, we will focus on examining within-context repeatability only. This approach allows us to assess the consistency of responses more accurately, in controlled repeatable settings, without the confounding variable of changing contexts.

(ii) After careful consideration, we have decided to retain the target age for testing. Whilst it is true that young birds of this age remain dependent on their parents in the wild, this dependency does not continue in captivity. By the time these hand-reared birds are about 5 days old and transferred to outdoor cages, they are already fully independent. Previous research from our laboratory on captive herring gulls reared under similar conditions (Troisi et al., in revision) and on wild-reared chicks from a neighbouring colony (Salas et al. 2022) shows, that strictly controlled testing of birds at this age provides ecologically relevant data on potential behavioural expression. Regarding the stability of neophobia at different ages, it is indeed recognised that this trait can change during development. However, the primary aim of our study is not to assess the stability of neophobia as a personality trait, but to investigate the effects of social context on neophobic responses.

(iii) While we acknowledge that our housing conditions do not mimic the natural group size of our study species, housing birds at their natural clutch size would require a more complex infrastructure. However, years of experience of releasing herring gulls reared under identical conditions and fitted with GPS

transmitters have shown highly comparable (social) behaviour to GPS-tagged wild individuals in a neighbouring colony. Nevertheless, we will explicitly discuss this relevant comment of the reviewer when discussing our results.

We declare that this revised Stage 1 RR remains original and unpublished. All authors approved the submission of the revised Stage 1 RR in its current form.

Comments Recommender

“In lines 98 to 102, you describe your predictions regarding reduced variance and reduced repeatability – please provide a full argument for the latter. “

As described in the first part of our rebuttal letter, which covers the main recurring issues, all predictions relating to repeatability have been removed.

“Please provide a full explanation as to the age at which you propose to test the gull chicks, and whether the results obtained from these tests can be considered stable across ages as well as how the relatively stableness or instableness refers to the overall framework of the study. “

For our reply to this question, please refer to the first part of our rebuttal letter, which addresses the main recurring issues.

”**In lines 135 to 136, you explain that under 10% of chicks may end up being from a different species (see also questions regarding this by one of the reviewers) and that thus you will conduct the analysis once with all birds tested, and once with the birds from the other species removed. The reviewer asks why you do not test the chicks at a time when you can differentiate the two species. Depending on your reply to this (and this also relates to the point I raised just above), please also explain why you you propose to conduct two analyses rather than just one, i.e. at the time when you know which birds belong to which species, (if you propose to keep both analyses) how you will deal with the results from both analyses, as well as whether the 10% ‘drop out’ has been taken into account in the sensitivity analysis and if applicable, change the sensitivity analysis to account for a 10% drop-out (alternatively, is an increase in the sample size to account for the 10% ‘drop out’ possible?). “**

We thank you and the reviewer for the very valid comment on the unconventional approach of performing the statistical analysis twice (once with and once without lesser black-backed gulls). On the basis of this comment, we have reconsidered our approach: we will perform the tests on all gulls (necessary for the target group sizes), but remove the results from the lesser black-backed gulls before performing the statistical analysis. Thus, only one analysis will be performed. We have also updated our power analysis to correctly account for a potential 10% drop-out rate. We have adjusted the manuscript as follows: “As gulls are reared from the egg, in a small number of cases (typically less than 10%), herring gull eggs are mistaken for those of the phylogenetically and ecologically related lesser black-backed gull. The species can only be determined after testing (when the individuals are older). Test data from lesser black-backed gulls (if any) will be excluded from subsequent analysis. We conducted a power analysis that accounts for

a potential 10% drop-out to ensure that even with this potential reduction, our study would still have sufficient statistical power (Cohen's f effect size of 0.17) to detect significant effects."

"In lines 204 to 205, you refer to the procedure regarding inter-rater reliability. Regarding this, do you have videos (or could you pilot this) that you could code and conduct inter-rater reliability to decide on the individual behaviours and the way in which these will be coded beforehand, i.e. NOT on the videos involving data of the main study? "

In the study by Troisi et al. (in revision), we coded similar behaviours in gulls, including "Start of trial", "Test arena entry", and "Eating". Inter-coder reliability was assessed and showed nearly perfect agreement, with an interclass correlation coefficient (ICC) greater than 0.90 for each behaviour.

Comments Reviewer 1

"My main concern is about the timing of testing with respect to the age of the birds and their upbringing. Birds will be hand-reared, which is fine but requires more justification and consideration how it might affect outcomes. More importantly birds will be raised in groups of 10 birds and moved into aviaries while in their nestling stage. Both seem to be highly unnatural as clutch sizes of herring gulls on average do not exceed three eggs and birds would normally remain in their nest or nearby when on roof tops until 45-50 days old. Particularly the large group size in relation to natural conditions might affect behaviours and specifically neophobia (birds might be more neophobic in the single situation as they are not used to it, whereas in wild birds often only one nestling survives). Furthermore, testing is planned around day 30 which is during the nestling stage (herring gulls fledge around 45-50 days of age). I cannot see the biological relevance testing gulls at a stage when they would not encounter novelty. Actually, they would still be fed by their parents and might still be supplementary fed by the caretakers in the current experiment. More importantly, I would not be able to make any predictions how the birds would respond as it does not reflect any natural situation. To test the predictions made in the introduction, birds should be tested after becoming independent and when they are in the flight cages. This ensures that birds are feeding on their own (i.e., reducing variance within individuals due to maturation) and also allows testing predictions at a stage when they would encounter novelty. This might require additionally habituating birds to the testing arena but would be worth the effort. "

Please see above for a detailed response to this comment.

"Furthermore, entire nests seem to be collected. However, the authors never mention any consideration of relatedness (i.e., keeping track of nest ID or only using one chick per nest). This needs consideration as relatedness affects responses (more related = more similar)."

As noted elsewhere in this rebuttal letter, we fully agree with the comment that this consideration was missing from the original design of the experiment. We therefore follow the recommendation that siblings be kept in separate groups during testing.

"The authors have considered inter-observer scores and also have control conditions. However, the experimental design could be improved regarding the length of testing. Currently, experiments last 10 minutes with the variable of main interest measuring the time between leaving the start box and feeding. Birds will leave the start box at different times providing different amounts of time to feed (less time when they leave the start box later). As 10 minutes seem to be already quite short, it would be better to start the timer once the bird has left the start box to give all birds 10 minutes to feed. The time to leave the start box can be restricted to 10 minutes. Both would result in values of 600 seconds in case birds do not leave or do not feed within the 10 minutes, which would be much better to handle than artificially assigning 600 seconds in case the birds do not feed within their remaining time after exiting the start box. "

We have taken full account of your comment by revising the task duration, which is now described in the manuscript as follows: "The first 10 min start when the door starts moving, the second 10 min start when all individuals have left the start box. The test session ends when all birds have interacted with the food or when 10 minutes have elapsed."

"Line 111pp: A) Why will juvenile gulls be used? Juveniles often respond differently to novelty than adults due to their lower experience. The use of juveniles should be better justified and background information provided about age differences in relation to neophobia and if possible in relation to social contexts. B) Furthermore, the authors mention control (food without novel objects) and experimental trials (food with novel objects). From the introduction it is unclear why this approach has been chosen. It would be nice to see a bit more about this approach. For example, many studies calculate the difference between feeding without and with the novel object or include both measures to control for differences in normal feeding latencies. It might also be mentioned that the difference between the two measures is a very good reflection of the actual fear invoked by the novel object (which is not always the case when the time to feed with object is taken without consideration of the time to feed without object). A paragraph critically evaluating both approaches (considering feeding times without objects or not) would help to understand and justify the own approach. C) Finally, the authors plan to use hand-raised birds. Again, more background information should be provided how origin (wild, hand-raised, captive bred) might affect responses. For example, being hand-raised or wild born has been shown to have substantial effects on neophobia responses (Neophobia Threshold hypothesis, Dangerous Niche hypothesis; Greenberg R (2003) The role of neophobia and neophilia in the development of innovative behaviour of birds. In: Reader SN, Laland KN, editors. Animal innovation. New York, Oxford: Oxford University Press. pp. 175-196; Greenberg R (1990) Feeding neophobia and ecological plasticity: A test of the hypothesis with captive sparrows. Anim Behav 39: 375-379). "

For a detailed response to these comments, please refer to the first part of our rebuttal letter, which addresses the main recurring issues, as well to similar comments below.

“Lines 139pp: It seems entire nests are collected. How is relatedness considered in testing and analyses? Individual identification is mentioned further down, but does this include nest ID? Closely related individuals are likely to respond more similarly, which can considerably affect results. Ideally, only one egg per nest should be used to have independence of data. In case, this is not possible, nest ID has to be considered in any analyses.”

See also next comments. We will implement the recommendation to assign siblings to separate groups during testing, and we will also include Nest_ID as a random effect in the analyses. However, if the variance between nests turns out to be small, we will remove this random effect. This will make our model less complex and will make it easier to interpret the other variances. Thus, the data will indicate whether or not including the nest-level is necessary.

“Lines 148-151: Will nestmates be assigned to different groups? This would be ideal. Why are 10 birds combined? This seems to be a quite unnatural group size for the nestling stage.”

See next comment/reply.

“Lines 153-155: Will the same 10 birds from the rearing stay together or will they be assigned to new groups? Again, having nestmates in different groups would be important. What about sex? Sex can affect neophobia. Will you sex birds and assign them into groups considering sex? With five days old, gulls would still be in the nest. How are birds kept until they fledge? Are they basically on the ground? This setup seems to be a major deviation from what the birds would experience in the wild (regarding number of birds and rearing conditions). Given that you are interested in social effects, this does not seem to be a good idea. Ideally, you should replicate wild conditions regarding number of birds together and rearing condition (i.e., keeping them in nest-like conditions until they fledge). “

Birds are reared together in groups of 10, which remain stable until after testing. We will implement the recommendation to assign siblings to separate groups, for which we thank the referee, during testing. However, for logistical reasons, it is not possible to determine the sex of the chicks prior to group assignment. Although we will explicitly account for possible sex differences in our statistical analyses, we do not expect an effect of sex, as herring gulls do not reach sexual maturity until they are 4 years old. While we acknowledge that our housing conditions do not mimic natural group sizes of our study species, housing birds at their natural clutch size would require a more complex infrastructure. However, years of experience of releasing herring gulls reared under identical conditions and fitted with GPS transmitters have shown highly comparable (social) behaviour to GPS-tagged wild individuals in a neighbouring colony. Nevertheless, we will explicitly discuss this relevant comment of the reviewer when discussing our results.

“Lines 157-159: Why do you test the gulls before they fledge? The first time the birds would encounter novelty in the wild is after they have fledged (although gulls nesting on roofs might wander around on the roof before fledging and experience some novelty). More importantly, as they are fed by their parents in the wild, they would not encounter novelty next to their food as it is directly transferred from the parent to the young. While it might be easier to test the birds before fledging, I cannot see the biological relevance and actually would not know what to expect/predict. A

better approach would be to move the fledged birds into their flight aviaries and get them used to feeding alone. Once this has happened tests can commence. This approach would have much more biological validity and would be comparable to other studies. "

For our reply to this question, please refer to the first part of our rebuttal letter, which addresses this issue in detail.

"Lines 162-164: Birds are allocated into two groups of five in their home cage. How is this done? Are they physically separated and if so, how is this done? With random allocation, how do you consider nestmates and sex?"

Birds are housed in groups of 10, with the group physically divided into two separate groups of five only during testing, using different pens (see Figure 2). As mentioned above, we follow the suggestion to divide siblings between the two groups. We have carefully considered sexing prior to grouping, but this is not feasible. Therefore, there may be an imbalance between the sexes within the experimental group (see also above).

"Lines 164-165: You mention a control object. This approach has never been mentioned before. Why do you use a control object rather than the feeding dish alone as control? Please justify your choice. Also, where will the control object be placed? Next to the food or somewhere in the enclosure? In the latter case it would still elicit neophobic reactions as the move of the object constitutes a change in the environment. "

We have clarified the use of a control object in the manuscript, with appropriate reference, as follows: "Conversely, in the 'control object' condition, a familiar object is placed in the home enclosure for six days prior to testing. By placing a familiar object behind the food plate prior to testing, we can observe responses during testing that are elicited by the novelty of the object and not just the presence of the object itself (see e.g. Greggor et al. (2015) for justification). Throughout the testing period, the familiar object remains in place and the novel object is introduced only during the testing sessions to avoid dishabituation from the familiar object."

"Tables and Figures: Have the table name and description above the table. I think you mention table 2 before table 1. Please revers the naming that table 1 is mentioned before table 2. For figures have the title and description below the figure."

We have adjusted the positions, references and legends of the tables and figures accordingly.

"Lines 178-179: Earlier it is mentioned that all birds are marked individually with colour ring combinations. Why is there additional marking necessary and how does this marking look like? "

As the videos are recorded with ceiling-mounted cameras, individual colour rings are not visible. Therefore, during group testing, each individual will be identified by a unique marker that will be visible in the top-view videos. Prior to conducting the tests, we will pilot methods to ensure that individuals are distinguishable in the group setting. The unique markers will be applied a few days in advance to allow the animals to become accustomed to them, and will be removed when the animals are transferred to the large flight cage. We have clarified this aspect in the main manuscript, as follows: "In order to distinguish the birds when they are being tested in a group, each individual will receive a unique marker a few days before the test, which can be easily detected by a roof-mounted camera, as the colour rings are not visible in the video recordings."

"Lines 184-186: Test conditions will last for 10 minutes only. The authors mention several studies having similar durations, but they are not on gulls. Have you done any experiments confirming that 10 minutes are enough for gulls? Birds often wait hours to approach food when a novel object is present. "

For a more detailed response regarding the duration of our experiment, please refer to the first part of our rebuttal letter, which addresses the main recurring issues. In short, we believe that this duration is appropriate, as 10 minutes was found to be sufficient to detect responses to a novel object in a study of neophilia in wild herring gulls (Inzani et al. 2023).

"Lines 191-192: How many birds are moved in the group condition? How do you assess latency to feed in the group control and experimental condition? Do you have a focal bird and record its time or do you record latency to feed for all individuals? Do you record order of feeding to account for other birds feeding first and therefore affecting responses in birds feeding later?"

Each group consists of 5 birds. Latency to feed will be determined by assessing the time between each bird leaving the start box and initiation of feeding. All tests will be videotaped, allowing us to follow the behaviour of each bird individually throughout the test session. The order of feeding is not included in the analysis. Individuals feeding first are not necessarily the first to approach the novel object and/or lead the group. While we acknowledge that it would be interesting to include group dynamics, this is beyond the scope of the study.

"Lines 202-204: Birds will have different times to leave the start box. This means that birds differ in the time available to feed from the food. As 10 minutes is already very short, it might be better to start the 10 minutes once the focal bird has left the start box to give all birds 10 minutes. In case, a bird does not leave the start box, one could terminate this phase at 10 minutes. These changes would make coding easier in case the birds do not leave the start box or do not feed. Alternatively, when you are interested in the time between making the food available and feeding then it would be better to measure the time between start of the experiment and the time to feed."

Please refer to the first part of our rebuttal letter which addresses the test duration. In short, we revised the duration of the task, following the suggestion that the 10-minute period should start as soon as the last bird

enters the test arena, giving each bird 10 minutes to eat. We will also give them 10 minutes in the start area.

“Lines 211pp: Will individual IDs be considered in one or the other way to calculate repeatability of responses on the individual level? How will you consider responses of other individuals?”

As described in the first part of our rebuttal letter, which covers the main recurring issues, all predictions relating to repeatability have been removed.

Comments Reviewer 2

“Page 4. L121, L133-136. The Authors will test 80 individuals of two species, as at the time of testing, the species cannot be ascertained yet. This may create problems in the group compositions, even if only 10% of the total number of individuals happen to be lesser black-backed gulls. Why not simply testing the animals later (i.e., before their release in weeks 8-10, see L159) if the species can be determined accurately then? “

For our reply to this question, please refer to the first part of our rebuttal letter, which addresses the main recurring issues.

“Page 5. Task design. Is placing a control object for three days in the home enclosure enough for the animals to habituate to it, and thus consider it as a control object? If possible, I would suggest having a control object for longer time in the home enclosure, especially as in some groups the control trials will be done immediately after the three habituation days, but in other groups some days later (which may lead to animals getting dishabituated in the meantime). “

We agree with the suggestion to extend the exposure to the control object. Therefore, we have revised the schedule and will now expose the gulls to the control object 6 days in advance. In addition, we expect minimal dishabituation, as the novel object will only be introduced during the test sessions, while the control object will remain in the enclosure between test sessions. We have modified the manuscript accordingly, please see comment above.

“Page 5. Objects. As two proposed objects are of the same colour (blue), would it be possible to exchange them for, e.g., an orange object (i.e. as this colour is also present in the multi-coloured ball)? Objects 3 and 4 seem to be objects that are commonly used for cleaning – can the Authors be sure that animals will consider them equally novel as a blue folder or a multi-coloured ball?”

This is indeed a very valid point and we have now replaced the blue folder with a white (rather than orange) one. The revised set of novel objects ensures that no two objects are the same colour (except for the multi-coloured ball). In selecting the novel objects, we have taken care to ensure that they are not familiar to gulls (including those commonly used for cleaning).

“Page 6, L188. Why placing a food bowl in front, and not on the side of the novel/control object? What kind of food and how much of it will be provided in the food bowl?”

We want to rule out possible directional preferences (left or right) and therefore place the food plate directly in front of the novel object. We have included further details in our main manuscript: “Next, a stacked plate of fish and an object (novel or control, depending on the condition) will be placed at the back of the enclosure, with the food plate placed in front of the object to rule out directional preference.”

“Table 1. Is the “additional buffer” included in the “zone of interest”? It is not clear to me how/if the “additional buffer” will be coded.”

We have clarified this aspect and rephrased it in the main manuscript as follows: “The 'zone of interest' is defined as a fixed rectangle encompassing the object and the food plate. To ensure comprehensive observation coverage, this area is expanded by the approximate body length of a 4-week-old gull (30 cm). This ensures that all relevant activities within and around the novel object are captured.

Table 1. Perhaps you can include in the “test arena entry” description that both feet need to be outside the start area AND that the bird needs to be within the test arena - I think that this would make the variable more comparable between the birds.”

Thank you for this suggestion. We have made the necessary changes to the ethogram to provide full clarity, and "test arena entry" will be coded according to the suggested criterion.

“Page 7. Video coding. Does the trial end when the bird starts eating, or is every trial 10 minutes long? If it is the latter, how will the Authors treat variables if there are several events of eating? In this vein, will you measure all times that the animals spend within the zone of interest, or just how long the animal spent there the first time (i.e. before it first started eating)? “

Trials end when the bird eats from the food plate. We encode the duration time of each period that a bird is in the zone of interest until the trial ends (i.e. the bird eats), and then calculate the total time in that zone by summing. We have included this suggestion in the main manuscript: “The first 10 min start when the door starts moving, the second 10 min start when all individuals have left the start box. The test session ends when all birds have interacted with the food or when 10 minutes have elapsed.”

“Page 7, Statistical analyses. Previously in line 96-97 it is stated that repeatability will be measured, but there are no details on the specific analyses/tests that will be done. It is also stated that the variance between individual and group trials will be compared, but there are no details about the proposed analyses.”

We have removed the repeatability prediction (please see above) and have now provided more details of the variance analysis in the text: “To specifically assess the variability in latency across individual and

group trials, we will compare the estimated variance components within our mixed-effects model. Variance for individual trials will be estimated from the Indiv_Dummy effect at the BirdID level. For group trials, the combined estimated variances of the Group_Dummy effect at both the BirdID and GroupID levels will be evaluated. This comparison aims to determine whether individual differences are more pronounced in solitary compared to group settings, with an expectation that individual variances and the total variance might be higher in individual trials. Additionally, an analysis at the BirdID level between the estimated variances of the Indiv_Dummy and Group_Dummy effects will further elucidate how individual differences manifest under different trial conditions, potentially highlighting the influence of group dynamics on individual behaviour."

"Page 14-15. "Rationale" should include more information about the power analysis, and "Sampling plan" part should include more information on the number and composition of the tested individuals and groups. For instance, as all groups will include 5 individuals, sexes will not be balanced, and perhaps some groups will include more than one lesser black-backed gulls. Do the Authors have a plan how to deal with these? Further, more details should be given regarding the test design (how the novel object trial versus control trial will be done, that the tests will be randomized, etc.)"

We have updated the table accordingly:
Sampling plan: "We will test 80 herring gulls twice across a 2x2 design. These four distinct conditions are: individual or group tests paired with a control or novel object. Each condition will be repeated twice. In the 'novel object' condition, birds are exposed to a pseudo-randomly selected novel object. Conversely, the 'control object' condition involves a familiar object, previously placed in their home enclosure for six days before testing. Testing trials will be randomised, see Supplementary table 1 in the main manuscript for a detailed testing schedule. Testing groups comprise 5 individuals by semi-randomly allocating gulls to one group. We will split nest mates across groups. Sexing is unfeasible prior to testing. While we will consider sex differences in our statistical analyses, we do not expect an effect of sex since herring gulls only reach sexual maturity at 4-years of age. Groups may also include a lesser black-backed gull. We will include all gulls for testing but will remove the lesser black-backed gulls prior to conducting the statistical analysis."

Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis:
"A-priori power sensitivity analyses were conducted in G*Power (Erdfelder et al., 2009), using a MANOVA. This indicated that our sample size of 80 animals is sufficient to detect a small effect of *Context, Group* and *Trial*. However, we will analyse our data with (G)LMMs, which are currently not covered by G*Power or most other power-estimation tools. These models are more flexible in assigning variance as they allow for the specification of both fixed and random effects. However, by accounting for unexplained variance, our proposed mixed-effect models are more powerful than the fixed-effect MANOVAs used in our sensitivity analyses."

"Page 5, lines 155-156. "depending on weather conditions" -> please note what kind of weather conditions are supplemented with heating plates?"

We have specified the weather conditions in the manuscript: "Outside, heating plates are provided during the first few days when night-time temperatures are forecast to drop below 5°C, or in the event of adverse weather conditions such as heavy rain or storms."

“Page 1. Abstract should include other expected outcomes of the tests, apart from the reduced variance (that are stated elsewhere in the text).”

We agree that it was confusing not to include all predictions in the abstract. However, since we removed the repeatability prediction (see main letter), the original abstract already included all remaining predictions, so the proposed change was no longer necessary.

“Figure 1. I would suggest to add the name of the respective hypothesis (instead of A, B and C scenario). “

We have implemented the suggested change.

“Page 6, line 192 “and a new test begins” -> do you mean that a new test for (a) new bird(s) begins?“

We have adjusted the sentence to clarify that a new test involves either a new individual bird or a new group of birds. “The testing session ends once all birds interact with the food, or once 10 minutes have passed. Next, the tester moves the tested bird(s) to the post-testing holding pen and starts a new test with a new (group of) bird(s).”

“Page 14. “latency measures”, “latency types” – please keep a consistent labelling otherwise one gets an impression that in one case it is about different variables, and in another that it is about variable distribution. Briefly mention what approach is proposed by Snijders and Bosker 2012; “a priori” should be written in italic”

Thank you for pointing out the inconsistency. We have adjusted the sentence to: “Models will be fitted to the different latency measures, both separately and in combination.” We also added the approach suggested by Snijders and Boskers (2012): “For the combined analysis, we will use the approach proposed by Snijders and Boskers (2012), which allows for the simultaneous analysis of multiple dependent variables in the case of nested data structures, thereby considering within-group and between-group variance in latency measures”

“Page 15. “depending on the social mechanism at play” is a bit vague, would suggest to elaborate”

We agree that the sentence was vague, we adjusted it in the manuscript to: “Social context may either modulate the group mean, the variance, or both. The risk dilution hypothesis suggests that being in a group will reduce both the mean and the variance of neophobia. Conversely, the negotiation hypothesis predicts

an increase in mean neophobia but a decrease in within-group variance. The social conformity hypothesis predicts no change in mean neophobia, but a decrease in variance. The design of our study allows us to validate or refute each of these hypotheses."