Round #1

Dear Editor,
Dear Recommender

We would like to thank you for the opportunity to revise and resubmit our Stage 1 registered report (RR), entitled "Do Ecological Valid Stop Signals Aid Detour Performance? A Comparison of Four Bird Species."

We would like to thank you and the reviewers for your very helpful comments. We have carefully considered and responded to each question below and in the manuscript. The revised sections are highlighted in the Stage 1 RR manuscript (which will be uploaded alongside this document on the PCI portal).

We have made several major changes to our Stage 1 RR. Most importantly, we defined our dependent variable 'persistence' much better (i.e., as the cumulative time spent in the species-specific 'barrier zone of interest'), clarified the 'multi-baseline' covariate (i.e., the variable provides a baseline for an individual's general motivational state), incorporated clear pre- and mid-test exclusion criteria (i.e., to avoid data skewness and to guarantee that individual experience with each barrier orientation is standardized), proposed more powerful statistical tests (i.e., to reduce unexplained variance of our model and increase the precision of our effects of interest) and clarified or justified aspects of the methodology (including the non-feeding period). To address the statistical issues raised by the reviewers, we have also consulted a statistical expert, who helped us with further fine-tuning our analysis protocol and selecting the most suitable models and approaches.

We would like to disclose again that we aim to test four different avian species. For two of those species we are restricted to the breeding season of the birds. This implies that testing will take at least a year. Furthermore, our planned start date (15-06-2023, as indicated on the PCI portal) is determined by the timing of the breeding season of the herring gull and is hence fixed.

We declare that this revised Stage 1 RR remains original and unpublished. All authors have approved the submission of the revised Stage 1 RR in its current form. I will be responsible for keeping my co-authors informed of our progress throughout the further editorial review process.

We would like to thank you for your time and effort, and for your consideration of our revised Stage 1 RR. Additional revision points raised by the recommender (but not addressed by the reviewers) are answered below.

Sincerely,
Anneleen Dewulf (on behalf of all authors)

Additional comments Recommender

"[…] I also noticed that the planned photoperiod of the different species is not the same"

That is correct. Unfortunately, we cannot control the photoperiod during testing of the herring gull (housed outdoors during the second half of June 2023, in accordance with the default policy of the Wildlife Rescue Center, WRC) nor can we adjust the photoperiod during testing of the canary (housed at indoor aviaries during late spring 2024, under natural L:D cycles, in accordance with the default policy of our collaborators at Antwerp University). In order to make the photoperiod of our species (more) comparable, Japanese quails and white leghorn chickens will housed under a L:D regime of 14:10 instead of 12:12. We have changed this in the manuscript:

> Subjects and Housing: White leghorn chickens and Japanese Quails
> "Once hatched, chicks will receive a unique colour ring combination prior to being housed in groups of 10 chicks per indoor enclosure ([…] photoperiod = 14:10 L:D)" (line number: 235-237).

Even though we cannot perfectly control for photoperiod, we think it is unlikely that this will influence our main results. First, in their large species comparison (N = 36 species), Maclean and colleagues (2014) did not find an association between day journey length and performance in the detour cylinder task or scores on another response-inhibition task. Second, we are not looking at absolute differences between species in our own study; instead, we are predicting an interaction between species and barrier type. It seems highly unlikely that such an interaction could be caused by (species) differences in photoperiod.

"First, I think it would be good if you could expand on the explanations of the relationships you expect between the covariates you chose to include (barrier neophobia and barrier order) and the outcome variables. I am worried that these two measures, rather than being covariates, might present as confounds."

We agree that the discussion of the relationship between the included covariates and other predictors and/or outcome variables was somewhat confusing. We have clarified this in the manuscript.

Additionally, the name choice of our baseline measure "barrier neophobia" was somewhat misleading. After all, the variable provides a baseline for a combination of (non-cognitive) motivational factors/traits, such as barrier or test box neophobia, but also food motivation, motivation to explore the environment, etc. We have therefore changed the description of our renamed 'multi-baseline' measure accordingly. Note that it is not our aim to distinguish between these various motivational factors but rather include a general measure for motivation. We explain this in the manuscript as follows:

> Introduction
> "[…] Fourth, non- cognitive, motivational states can influence detour performance (Kabadayi, Krasheninnikova, et al., 2017; Van Horik, Langley, et al., 2018). Therefore, we will collect for each individual a 'multi-baseline' measure of their general motivational state (which could be a combination of, e.g., non-transparent

obstacle neophobia, test box neophobia, food motivation, motivation to explore). […] We will include this as a covariate in our statistical models to increase the likelihood of detecting barrier type effects within species conditional on/adjusted for the 'multi-baseline' measure of an individual's general motivational state." (line number: 153-160).

With regards to the barrier order, we wrote following in the revised report:

Statistical Analysis

"[…] In addition, we will add two extra explanatory variables to the model: […] and *Barrier Order* (with two levels: did the individual receive the horizontal-bar barrier on the first test day 1 and the vertical-bar barrier on the second test day; or vice versa), as species might demonstrate superior performance with the last encountered barrier, irrespective of its type and ecological validity)" (line number: 432-438).

Second, in the study design template, I think you could split some of the information by the three questions you set up. I recognise that it is difficult to clearly separate between these points, but I think the (i) 'Analysis Plan' and (ii) 'Theory that could be shown' columns could have separate entries for each question that focus on (i) the specific effects of the statistical analysis and (ii) the specific rationale underlying each question.

We have revised and clarified the columns "Analysis plan" and "Theory that could be shown wrong by the outcomes" of the study design template.

Analysis Plan:

Question 1:

A (G)LMM with type III sum of squares will be used to analyze our two dependent variables, namely 1) the latency to detour and the 2) cumulative time spent in the species-specific 'barrier zone of interest'(persisting).

Both models will include the between-species factor: *Species* (i.e., 4 levels) and both within-species factors: *Barrier* (i.e., 2 levels) and *Trial* (i.e., 3 levels), and the two control variables (as covariates), namely (a) a *'multi-baseline'* measure of an individual's motivational state (and its interaction with *Species*, as we will mean-center this *'multi-baseline'* measure within *Species*), and (b) *Barrier Order* (i.e., 2 levels). Individual birds and enclosure (social group) will be included as random effects in the models, with individual birds nested in enclosures. In addition, we will include by-individual (nested in enclosures) random slopes that can vary for the levels of *Species* (corresponding with species-specific intercepts).

Model plots will be generated by means of the package *performance* (Lüdecke et al., 2021) to inspect for violations of the model assumptions: 1) heteroscedasticity (plotting the square root of the residuals (y-axis) and fitted values (x-axis)), 2) non-normality of residuals (plotting the sample quantiles (y-axis) on the standard normal distribution quantiles), and 3) outliers (plotting standard residuals (y-axis) and leverage). Additionally, the multicollinearity between fixed main factors (via the variance inflation factor, VIF) and the autocorrelation between residuals (via a Durbin-Watson-Test) will be calculated via functions provided by the *performance* package (Lüdecke et al., 2021). Potential violations of model assumptions will be addressed by transforming the (in)dependent variables (i.e., via log-transformation) or by changing the error distribution (family) or the link function of the model (switching a default LMM that will be fitted to a GLMM). Fixed main effects with a VIF of >5 will be removed and logical outliers (i.e., recording/entry errors) will be inspected and corrected (if possible).

In the case that the outlier cannot be corrected, all data of that individual will be excluded from all statistical analyses.

In case we find (a) significant *Barrier* x *Species* interaction - effect(s) **(Question 1)** further post-hoc Bonferroni-Holm corrected linear contrasts upon the model will be performed to compare performance with different (ecological valid) barriers per species **(1:1, 1:2, 1:3)**.

Question 2:
In case we find (a) significant main effect of *Trial* **(Question 2)** further post-hoc Bonferroni-Holm corrected linear contrasts upon the model will be performed to compare performance over trials **(2:1).**

Question 3:
In case we find (a) significant three-way *Species* x *Barrier* x *Trial* interaction effect(s) **(Question 3 explorative)**, further exploratory Bonferroni-Holm corrected linear contrasts upon the model will be performed to compare *Trial* performance of *Species* on different types of the *Barrier* **(3:1 explorative).**

Theory that could be shown:
Question 1:
We propose that stop-signal detection (hence, barrier detection) is a crucial cognitive building block of RI across species (Verbruggen et al., 2014), including birds.
Here we will take this idea one step further and propose that ecologically valid signals are easier to detect (or to perceive as a stop signal) and this will enhance stopping.
The role of stop-signal detection in avian response inhibition, and in particular, the interaction with the ecological niche of the species, should be revised if we cannot replicate the previous work (Regolin et al., 1994; Zucca et al., 2005)

Question 2:
We propose that detour performance improves over trials. Extensive work on skill acquisition in humans has shown that performance generally improves rapidly at first and then more slowly over time (see e.g., Logan, 1988, Thorndike, 1913). If we do not find a difference between trials, this would indicate that detouring cannot be learned easily by avian species.

Question 3:
We will *explore* if the learning effect (i.e., improved detour performance across trials) will depend on the ecological validity of signals.
If we do not find such a three-way interaction effect, we can conclude (a) that superior detour performance with ecological valid than non-valid trials is independent of trial number (in case we do find a *Species* x *Barrier* interaction effect) or (b) that the interaction between the stop signal and the ecological niche of the species should be revised (in case we do not find a *Species* x *Barrier* interaction effect).

Comments Reviewer 1

It was a pleasure to review this well-written and very detailed pre-registered report. The data analysis plan is relatively sound, and the methods are well-designed and appropriate to answer the question at hand.

I have a few questions and potential recommendations regarding the proposed behavioural parameters (1 and 2), the rationale for the statistical analysis (3), and ethical issues arising due to food deprivation (4).

1. Parameters for behavioural analysis: I am a bit critical of using "touches barrier with beak" as indicator of persistence as physical inspection/pecking might differ across species. This would run the risk of zero-inflated data for some species (i.e., those with a low motivation to establish physical contact with the barrier) – hampering comparisons between species, and also barrier-design. Unless the background literature would indicate otherwise, I would argue that proximity to the barrier (or certain areas of the barrier) would probably be a better indicator for persistence.

We agree that the proximity to the barrier (which we define in the revised manuscript as the cumulative time spent in the species-specific 'barrier zone of interest') is a better indicator for persistence. We have therefore added this information to our manuscript:

> Introduction
> "Fifth, our study will consider […] the time spent in proximity to the barrier (persistence)." (line number: 160-162)
>
> Video Recording and Analysis
> "[…] Second, the time spent persisting (in seconds) will be calculated as the cumulative time that the individual spends in the species-specific 'barrier zone of interest' (size = Barrier L x 25% of the Barrier-Entry Distance; L x W, see table 2 for the species-specific dimensions)." (line number: 388-391)
>
> "**Persisting:** At least the bird's whole head crosses the (fictitious) lines of the rectangular-shaped, species-specific 'barrier zone of interest'." (**Table 3**, the description of the behaviours that will be coded in BORIS.)

In addition, we have added a figure to the manuscript that will help with the interpretation and visualization of the to-be-coded events in BORIS, including the proximity to the barrier, which will be used to calculate the new measurement of persistence (**Figure 3**, page 12)

2. Inclusion of maximum trial time and consideration of additional behavioural parameters: The authors set the maximum trial limit to 135s and opt to include trials in which the barrier is not being detoured as data points with "135". This can be problematic for at least 2 reasons: 1. If subjects do not detour the barrier in a considerable number of trials, this will skew the data and might hamper further statistical analysis. 2. Failing to detour the barrier can be due to an inability to find a detour route but can also be caused by other factors such as lack of motivation, increased stress, or distraction (i.e., subjects that might otherwise be easily able to detour will get assigned with the highest value). To get around 1., it might be advantageous to also score accuracy (yes/no) and analyze it in a corresponding GLMM (similar to what has been done in previous studies using this paradigm).

We agree that (a) data skewness could be an issue for the analysis and interpretation of the findings and (b) the individual's motivational state (e.g., a combination of food motivation, stress, distraction or over exploration) can influence (the lack of) detour performance on the current/subsequent trials.

Both issues are related, and based on your comments, we have reconsidered the inclusion/exclusion criteria of individuals. We agree that including (a) individuals that are demotivated or distressed and (b) individuals that are distracted and overly explorative, can be problematic as it might hamper drawing (firm) conclusions on the within-species level, especially if maximum scores are assigned. We have therefore adjusted the exclusion criteria of our manuscript. By excluding the 'non-participating' individuals (based on performance in the sessions), we also think it is less likely that the data will be highly skewed. In case that we still obtain (positively) skewed data, switching from a default LMM to GLMM (see our reply to comment 3) enables us to select an appropriate error distribution (e.g., gamma or inverse gaussian distribution family) capable of modelling this type of data, and hence will aid further statistical analysis.

> Data exclusion criteria
> "[…] Birds that did not detour around the barrier nor entered the species-specific 'barrier zone of interest' in a test trial will be excluded from subsequent test trials (and data of that individual will be excluded from all statistical analyses). This mid-test exclusion criterion will be applied for two reasons. First, birds that do not obtain a measure for one of the two dependent variables within 2 minutes are likely to be unmotivated or be in distress. Furthermore, observations from similar RI test paradigms in our lab demonstrate that such individuals are unlikely to eat at all with a prolonged test time or on subsequent test trials (within the same day) \footnote{In a continuous RI task with a sample size of 80 herring gulls, birds that failed on the first trial, were likely to fail again on the second trial of the same test day (Dewulf et al., (2022)}. In addition, removing birds from subsequent trials (rather than assigning a maximum trial limit for both dependent variables) reduces the risk of data skewing.
> Individuals that have left the species-specific 'test box zone of interest' (size = 2 times the Barrier-Entry Distance, see table 2 for the species-specific dimensions) without touching the food (bowl) will also be excluded from further testing and all analyses. This mid-test exclusion criterion assures that we avoid confusing general exploration behaviour (without initial interest in the food) with successful detour performance (which assumes interest in the food). Thus, by excluding birds with differential trial experiences (due to e.g., demotivation, distress, distraction or exploration; for a similar exclusion criterion see, Van Horik, Beardsworth, Laker, Langley, et al., 2019), we aim to ensure that each barrier orientation is standardized within- and between species.. Note that we expect that we can maintain our sample size by replacing all excluded birds, because we generally incubate 20% more eggs than the number of individuals required for the testing (to account for possible drop outs during the whole study). (line number: 401-419)

In addition, we have added a figure to the manuscript that will help with the interpretation and visualization of the to-be-coded events in BORIS, including leaving "the test box zone of interest" (**Figure 3**, p12).

We have opted <u>not</u> to include accuracy as a third dependent variable in our manuscript for three main reasons. First, accuracy is also captured by our variable "persisting" (line number:

163-165). We opted for the persistence variable as it is more informative; i.e., not only do we measure if the bird makes an error or not, we also measure for how long it persisted in this erroneous behaviour. Second (and related to the previous point), adding a binary response variable requires a substantially larger sample size to obtain reliable (well-powered) results. For example, an *a-priori* power sensitivity analysis for a logistic regression analysis (alpha corrected for multiple tests for each dependent variable =.0167, Power = .80, X distribution = binomial, odds ratio = 0.669, tails = 2) with G*Power (Faul et al., 2009) indicates that a sample size of 1048 animals per species is required to detect a main effect of barrier orientation within one species. Note that currently only logistic regression analyses with one predictor are included in G*Power. Yet, we believe that a generalized linear mixed-effect regression analysis (GLMM) with a binary response variable that captures the variance of the model better will not drastically decrease the sample to a size that is feasible given our aviary constraints. Third, running multiple tests (i.e., an additional GLMM for the dependent variable accuracy) increases the type 1 error due to multiple tests for each of the multiple dependent variables.

3. Statistical analysis: I was wondering why the authors opt for an ANCOVA rather than a (G)LMM – with the latter being more flexible in assigning variance / estimating effects. Mixed models are generally more powerful compared to conventional repeated measures AN(C)OVAs and they also have fewer assumptions (e.g., sphericity). In addition, the authors need to state how they will proceed if model assumptions for ANCOVA (or LMMs) cannot be met (e.g., the need for data transformation).

See also here: Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language, 59(4), 390-412.

Initially, we had opted for an AN(C)OVA as both Regolin et al., (1994) and Zucca et al., (2005) utilized a similar data analysis approach. However, we agree that (G)LMM's are more flexible in assigning variance and estimating effects (including, adding a random effect of social group), so we have changed our analysis plan accordingly (after further consultation with a statistical expert):

> Statistical Analysis
> "[…] Models will be fitted my means of the *lme4* package (Bates et al., 2015) and parameter estimation and p-values for the generated models will be provided by means of the *lmerTest* package (Kuznetsova et al., 2017) via the Satterthwaite's degrees of freedom method (linear mixed model, LMM) or via the *carData* (Fox, Weisberg, and Price, 2022) and *car* (Fox and Weisberg, 2019) packages (generalized linear mixed model, GLMM). For the (G)LMM, we will use partial eta-squared (η2p) as effect sizes and they will be calculated by means of the *effectsize* (Ben-Shachar et al., 2020) package.
>
> A (G)LMM with Type III sum of squares will performed on the latency to detour and the cumulative time spent in the species-specific 'barrier zone of interest' (persisting). Both models will include the between-species factor: *Species* (i.e., white leghorn chickens, Japanese quails, herring gulls and canaries) and both within-species factors: *Barrier* (i.e., vertical- and horizontal-bar) and *Trial* (i.e., 1-3), and their interactions. In addition, we will add two extra explanatory variables to the model: a *'multi-baseline'*

measure of an individual's general motivational state (and its interaction with Species, as we will mean-center this 'multi-baseline' measure within *Species*, see Chen et al., 2014 for an example of within-group centering); and *Barrier Order* (with two levels: did the individual receive the horizontal-bar barrier on the first test day 1 and the vertical-bar barrier on the second test day; or vice versa), as species might demonstrate superior performance with the last encountered barrier, irrespective of its type and ecological validity. Individual birds and enclosure (social group) will be included as a random intercept in the models, with individual birds nested in enclosures. In addition, we will include by-individual (nested in enclosures) random slopes that can vary for the levels of *Species* (corresponding with species-specific intercepts).

Model plots will be generated by means of the package *performance* (Lüdecke et al., 2021) to inspect for violations of the model assumptions: 1) heteroscedasticity (plotting the square root of the residuals (y-axis) and fitted values (x-axis)), 2) non-normality of residuals (plotting the sample quantiles (y-axis) on the standard normal distribution quantiles), and 3) outliers (plotting standard residuals (y-axis) and leverage). Additionally, the multicollinearity between fixed main factors (via the variance inflation factor, VIF) and the autocorrelation between residuals (via a Durbin-Watson-Test) will be calculated via functions provided by the *performance* package (Lüdecke et al., 2021). Potential violations of model assumptions will be addressed by transforming the (in)dependent variables (i.e., via log-transformation) or by changing the error distribution (family) or the link function of the model (switching a default LMM that will be fitted to a GLMM). Fixed main effects with a VIF of >5 will be removed and logical outliers (i.e., recording/entry errors) will be inspected and corrected (if possible). In the case that the outlier cannot be corrected, all data of that individual will be excluded from all statistical analyses." (line number: 421-453).

On OSF, we have created a folder "AnalysisPlan" which contains a R-script with the to-be-used model (assumption) functions, and specific packages needed per function.

Note that we will now conduct post-hoc linear contrasts (instead of paired-t-tests, as was the case in the first draft of this Stage 1 RR), as our statistical advisor argued that performing post-hoc linear contrasts upon the model are more powerful (due to a more accurate estimation of the error).

4. Ethics: food deprivation (from 4:00PM - 8:30AM the following day) appears quite exhaustive for small birds (e.g., canaries) – is there literature showing that these are deprivation times that are commonly used and do not pose strong additional stress or harm to the animals? Otherwise, I would argue to reduce deprivation times considerably.

We have adjusted the period so that it is more in line with previous work (for the canaries, see e.g., Müller et al., 2008) or with the standard procedures at the Wildlife Rescue Centre (where most of our research is taking place). We have also further clarified this in the main manuscript:

Procedure

"[…] Food is provided *ad libitum*, but in the evening before an individual's habituation or testing day, the feeders will be removed from the enclosures at 6PM (after the last feeding time). This will create a non-feeding period during the night (which is normal and also happens in non-experimental conditions), followed by

(shortly) delayed feeding in the morning to prevent birds from overindulging prior to habituation or testing. This is in line with other studies using the same species (chicken: e.g., Bollweg and Sparber, 1998; quail: e.g., Ueno and Suzuki, 2014 and unpublished data from our lab; herring gulls: e.g., Dewulf et al., 2022; domestic canaries: e.g., Müller et al., 2008). After all individuals of one enclosure have completed the habituation or testing trials for the day, food will be again provided *ad libitum*." (line number: 324-331).

Note that we removed 'food deprivation' as this term was incorrect (i.e., non-feeding periods during the night are normal).

I also have some rather minor comments:

<u>Introduction</u>

5. 1st paragraph, first sentence: Needs reference.

We have added the reference to the manuscript.

"Response inhibition (RI) refers to stopping or cancelling actions that are no-longer relevant, inappropriate, or overly risky" (Verbruggen and Logan, 2008b, 2017)" (line number: 15-16).

6. 1st paragraph, last sentence: "if they fail to stop" – what does this mean? Stop what?

We have added the clarification to the manuscript.

"[…] the animals may be predated if they fail to stop foraging when a predator emerges" (line number: 23-24).

7. 2nd paragraph: it might be worth outlining some details about the limitations of detour tasks to assess response inhibition, e.g., Horik et al. 2018.

van Horik, J. O., Langley, E. J., Whiteside, M. A., Laker, P. R., Beardsworth, C. E., & Madden, J. R. (2018). Do detour tasks provide accurate assays of inhibitory control? Proceedings of the Royal Society B: Biological Sciences, 285(1875), 20180150.

We agree that performance on the detour task can be influenced by e.g., differential individual experiences with (transparent) obstacles (Van Horik, Langley, et al., 2018; Kabadayi, Krasheninnikova, et al., 2017), training treatments that enable acquiring detour motor routines (Van Horik, Beardsworth, Laker, Whiteside, et al., 2020), variations in non-cognitive traits (Van Horik, Langley, et al., 2018; Kabadayi, Krasheninnikova, et al., 2017) and cognitive development (Kabadayi, Jacobs, et al., 2017; Kabadayi, Krasheninnikova, et al., 2017; Verbruggen, McLaren, et al., 2014).

We have added these limitations of the detour task in our manuscript when discussing the methodological and conceptual shortcomings commonly raised in the detour literature (incl. the Regolin et al. (1994) and Zucca et al. (2005) study) and/or procedure. We did not incorporate this in the 2nd paragraph of the introduction to keep the focus on gaining better understanding on how perceptual processes contribute to response inhibition.

Introduction

"In our partial replication, we will make several changes to address commonly raised concerns in the detour literature (including concerns raised in the previous section, see table 1) […] Fourth, non-cognitive, motivational states can influence detour performance (Kabadayi, Krasheninnikova, et al., 2017; Van Horik, Langley, et al., 2018). Therefore, we will collect for each individual a 'multi-baseline' measure of their general motivational state (which could be a combination of, e.g., non-transparent obstacle neophobia, test box neophobia, food motivation, motivation to explore). This 'multi-baseline' measure will be obtained with an opaque barrier during habituation (see below). We will include this as a covariate in our statistical models to increase the likelihood of detecting barrier type effects within species conditional on/adjusted for the 'multi-baseline' measure of an individual's general motivational state. […] \footnote{[…] Second, performance in the detour task can be influenced by differential individual experiences with transparent obstacles (Kabadayi, Krasheninnikova, et al., 2017; Van Horik, Langley, et al., 2018).} Seventh, detour performance of the different species will be compared when they are on similar levels in their developmental trajectory (see e.g., Kabadayi, Jacobs, et al. (2017), Kabadayi, Krasheninnikova, et al. (2017) and Verbruggen, McLaren, et al. (2014) for the influence of cognitive maturation on RI), and again, with similar experiences in the enclosure, keeping in mind the precocial-altricial spectrum (see below)." (line number: 141-172).

Procedure:

The current habituation set-up (i.e., the food bowl in front of the barrier) is designed in such a way that acquiring a motor routine during habituation is unnecessary and cannot confound subsequent detour performance with the barred barriers (Van Horik, Beardsworth, Laker, Whiteside, et al., 2020). (line number: 338-340).

8. Last sentence of 2nd paragraph: other factors than unpredictability might affect the expression of diverse foraging strategies so this suggestion might be a bit too general. In addition, "learn" would imply that there will be a gradual increase over lifetime/experience – is this what the authors imply?

We have adjusted the information in the manuscript.

"[…] Combined, these findings suggest that RI development is facilitated in e.g., environments with high social demands or environments that promote the expression of diverse foraging strategies." (line number: 37-38).

9. 3rd paragraph, second sentence: please outline why "inhibitory control" would be even less suitable to serve as a general umbrella term ("or even worse…").

We have further clarified this. The term inhibitory control is well-suited to be a general umbrella term for response inhibition, but it is important to realize that different forms of inhibitory control might not be (mechanistically) related. For example, work within the neuroscience domain indicates that delaying gratification and inhibiting actions (both fall under the 'inhibitory control' umbrella) involve different neuro-cognitive mechanisms (e.g., different neurotransmitters and different neural networks are involved).

"[…]. Typically, performance in the detour task has been linked to the variation in the effectiveness of a single cognitive control function, 'response inhibition', or more

> generally, 'inhibitory control' (which is an umbrella term for various types of inhibition, which may or may not be related to each other; Bari and Robbins, 2013). However, by referring to general, ill-defined cognitive constructs such as RI (or even worse, a general umbrella term such as 'inhibitory control'), we do not explain the underlying cognitive mechanisms or building blocks of stopping (Verbruggen, McLaren, et al., 2014), as the explanation is 'just as mysterious as the thing it is supposed to explain' (Press et al., 2022)." (line number: 40-46).

10. 3rd paragraph, last third: "Furthermore, these core process …" – please explain what is meant with timescales here.

We have clarified the information in the manuscript.

> "[…] Furthermore, these core processes can be modulated by a set of processes that take place on shorter (seconds, minutes, hours or days) and longer (months or years) timescales" (line number: 50-52).

11. 5th paragraph, last third: "The authors found that RI performance (…) was worse …" what does "worse" refer to here? Longer or shorter latencies to perform the detour?

We have clarified these findings in the manuscript.

> "[…] The authors found that RI performance was impaired (i.e., the time required to successfully detour around the barrier was prolonged) when faced with vertical- than horizontal-bar barriers." (line number: 75-77).

12. Table 1: Without reading the full introduction it was not clear which species were included in Zucca (2005) – might be worth considering adding some additional lines as separators

We agree that the table 1 was hard to read, so we have adjusted the lay-out (see page 5 in the revised manuscript).

13. Why was this specific baseline measure for neophobia chosen? Please briefly elaborate and consider adding references.

The name choice of our baseline measure "barrier neophobia" was indeed confusing. After all, the variable provides a baseline for a combination of (non-cognitive) motivational factors/traits, such as barrier or test box neophobia, but also food motivation, motivation to explore the environment, etc. We have therefore changed the description of our renamed 'multi-baseline' measure accordingly. Note that it is not our aim to distinguish between these various motivational facets. Instead, we just wanted to include a general measure for motivation. We explain this in the manuscript as follows:

Introduction

"[…]) […] Fourth, non-cognitive, motivational states can influence detour performance (Kabadayi, Krasheninnikova, et al., 2017; Van Horik, Langley, et al., 2018). Therefore, we will collect for each individual a 'multi-baseline' measure of their general motivational state (which could be a combination of, e.g., non-transparent obstacle neophobia, test box neophobia, food motivation, motivation to explore). This 'multi-baseline' measure will be obtained with an opaque barrier during habituation (see below). We will include this as a covariate in our statistical models to increase the likelihood of detecting barrier type effects within species conditional on/adjusted for the 'multi-baseline' measure of an individual's general motivational state." (line number: 153-160).

Procedure

"[…] During the second and third habituation day, an opaque barrier will be placed just behind the coloured food bowl. This will allow us to obtain a 'multi-baseline' measure of an individual's general motivational state (which could be a combination of e.g., non-transparent obstacle neophobia, test box neophobia, food motivation, motivation to explore; see below). The current habituation set-up (i.e., the food bowl in front of the barrier) is designed in such a way that acquiring a motor routine during habituation is unnecessary and cannot confound subsequent detour performance with the barred barriers (Van Horik, Beardsworth, Laker, Whiteside, et al., 2020)." (line number: 334-340).

14. Figure 1: please reconsider using red/green colour differences – maybe change to other colours or shades of grey.

Figure 1 has been adapted to a shades of grey colour scheme (see page 6 in the revised manuscript).

Methods

15. Comparing birds of different ages: I fully see the need to account for the time birds can experience their enclosure (canaries vs all other species) but I was wondering from a developmental perspective whether the additional age might give canaries a general head start in the task (although a mere species comparison is not the aim of the study).

The canaries are indeed being tested when they are older (in terms of days). In theory, this could give them an advantage as stopping improves with cognitive (inhibitory) maturation (for reviews, see Bunge, Mackey, & Whitaker, 2009; Diamond, 2013). However, we have to take into account that developmental trajectories differ substantially between species. Canaries are altricial, i.e., they are initially blind (hatching day 1-8) and remain in the nest where they are fed by their parents (hatching day 1-25). In order to guarantee that they can solve the detour task, it is critical that they develop the needed perceptual and motoric skills (i.e., motor experience, coordination and sensorimotor experience; see e.g., Kabadayi et al., 2017).

It should also be noted that we make specific predictions about the *Species* x *Barrier* interaction. We deem it unlikely that such an interaction could be caused by age differences. Nevertheless, it is an issue we plan to comment on when interpreting the findings in the discussion section of our Stage 2 manuscript.

16. Assessing barrier neophobia: running only one trial makes it quite prone to outliers/distractions. Shouldn't it also be corrected with a different baseline, e.g., the time needed without the barrier (as some species might be simply slow in approaching the food in general)?

We agree that running only one trial can make our baseline measure prone to outliers. We have therefore changed our procedure and will take the average of habituation day 2-3 instead. As noted above, we have changed the name of our variable to a 'multi-baseline' measure. We have clarified both aspects in the manuscript.

> Procedure
> "[…] During the second and third habituation day, an opaque barrier will be placed just behind the coloured food bowl. This will allow us to obtain a 'multi-baseline' measure of an individual's general motivational state (which could be a combination of e.g., non-transparent obstacle neophobia, test box neophobia, food motivation, motivation to explore; see below). (line number: 334-338).
>
> Data Processing and Analysis
> "[…] Third, a 'multi-baseline' measure of an individual's general motivational state (in seconds) will be calculated, by *averaging* the time between leaving the start box and touching the food (bowl) placed in front of the opaque barrier on habituation trial 2 and 3. Note that if a bird is unsuccessful on trial 2, a non-averaged 'multi-baseline' score will be calculated based on habituation trial 3 only" (line number: 391-394).

17. Unpublished literature. Is there any chance to make Troisi et al. and Garcia-Co et al. publicly available (e.g., via preprinting) as they are used to justify the task measurements for two of the tested species?

On OSF, we have created a folder "figures", which includes the R-script and morphological unpublished datasets (i.e., Troisi et al., 2022 and Garcia-Co et al., 2022) to generate figure 2 of the manuscript (i.e., this figure visualizes the growth curve of tarsus length for each species). The species-specific rescaled apparatus size at test age can be derived (and justified) from these growth curve.

At this moment, the data are in a private OSF repository that can be accessed by the Managing Board, specialist recommender and both reviewers via the manuscript repository URL provided at submission. We will make all data publicly available at stage 2 RR acceptance.

Comments Reviewer 2

This registered report describes an experimental setup, to evaluate the influence of stop-signal detection in the performance of birds in a detour-barrier task, assessing whether results will be predicted according to the ecological niche of each species. In the proposed research plan an experimental procedure will compare the performance in inhibitory response of 4 different bird species, namely white leghorn chickens, Japanese quails, herring gulls and domestic canaries, all hatched and raised in captivity. The proposed research plan is also a partial replication of 2 previous studies on the same issue (Regolin et al. 1994, Zucca et al. 2005), but improving some critical aspects raised from those studies.

In my opinion the research questions are scientifically valid and personally I think they are interesting to research in cognition, specifically to better understand the roles and evolution of response inhibition in birds.

Overall, the research plan, the experimental setup, and the statistical procedure proposed sounds reasonable, plausible and logic to test the hypothesis presented, and able to drive robust results. The research plan and methodology also seem highly feasible and with enough detail to be understandable and replicated. Thus, I have no doubt that the research plan is possible to occur and give valuable results. Finally, I think the authors anticipated in a reasonable way the control conditions needed to validate results of the test procedures.

However, I think the research plan would benefit from some further clarifications in the methodology, especially explicitly justifying some methodological decision. I made comments for that below. I hope the authors find them useful.

From the critics the authors provide to the previous works of Rogolin et al. and Zucca et al. I agree that in this proposed plan the consistency in the sample size, with an enough sample size to test their prediction (as the authors show), is an improvement regarding the previous 2 studies. Furthermore, I also agree with the proposed adaptation of the test apparatus to the body size of each species and also with doing a simple detour barrier task to each species, in order to have more robust comparisons of the results between species. Finally, I also agree that using a more standardized reward such as food makes results between species more comparable.

18. However, I do not understand why only 3 trials will be performed per species. If authors want to take into account the influence of learning, are 3 trials enough to assess this goal? A previous comparative work by McLean et al. (2014, 10.1073/pnas.1323533111) and in several other empirical works with different species, the number of trials given to each individual is higher. In several of those studies with fixed number of trials, researcher have chosen around 10 trials to perform, or to do trials until a learning threshold is reached (i.e., a specific number of consecutive trials where individuals successfully detoured the barrier to retrieve food). I agree that in the case of this proposed work the same number of trials should be done to each species, but I lack to understand why only 3 trials will be made, especially considering that authors clearly intend to assess the influence of learning or performance improvement across trials.

We agree that more trials would be even better to assess learning. However, 10 trials per barrier type is not feasible (given that we aim to test 60 individuals of four different species). Importantly though, most learning generally takes places during the first trials (see e.g., Logan, 1988; Thorndike, 1913). This seems also the case in the detour task. For example, Van Horik, Langley, et al. (2018) found in pheasants (*Phasianus colchicus*) that the latency to obtain the reward in the detour task with a transparent barrier decreased from ≈ 150 s (trial 1), to ≈ 80 s (trial 2), followed by ≈ 70 s (trial 3). Additionally, the authors demonstrated that the pecks at the transparent barrier decreased from ≈ 90 pecks (trial 1), to ≈ 30 pecks (trial 2), followed by 5 pecks (trial 3). We have recently obtained very similar results in Japanese quails (Willcox et al., manuscript in preparation).

Extensive training with one barrier type would also increase the probability that we would observe order effects (horizontal vs. vertical-bar barriers first), complicating the interpretation of our findings.

For these two reasons, we deemed three trials appropriate to assess initial learning. Note that we have incorporated your comment (and the above reply) in the study design template, column "Theory that could be shown wrong by the outcomes" (see our reply to the additional comments of the recommender).

> "We propose that detour performance improves over trials. Extensive work on skill acquisition in humans has shown that performance generally improves rapidly at first and then more slowly over time (see e.g., Logan, 1988, Thorndike, 1913). If we do not find a difference between trials, this would indicate that detouring cannot be learned easily by avian species."

19. I think the methodological procedure should be clarified. It states that "3 trials per day" will be made, but it is not clear whether it would be overall or to each bird. Is each bird entering in habituation or testing days at a time, or this happens to some number of individuals each time or perhaps the group. It is not clear to me how this part of the procedure will occur.

We agree that the methodological procedure was somewhat confusing. We have clarified this in the manuscript. In addition, we have added a section about 'batch testing', which was not mentioned in the earlier version of this manuscript.

> <u>Procedure</u>
> "On the three habituation days (08:00 AM - 10:30AM), each bird will receive 1 trial per day where it can freely explore the test box and feed from a centrally placed coloured food bowl. […]" (line number: 333-334).
>
> "On the two testing days (10:30AM - 02:30PM), each bird will perform one session, each consisting of 3 trials with one barrier type. The order of barrier type (i.e., horizontal-bar or vertical-bar barrier) will be pseudo-randomized within and between species, across the two testing days." (line number: 341-343).
>
> "Due to the natural breeding season of the wild herring gull and the domestic canary, birds hatch non-simultaneously. In order to guarantee an appropriate test age (see above), we will group individuals of a similar age per enclosure; and then habituate or

test birds per enclosure (by taking into account the average age of the enclosure). Although there is no fixed breeding season for Japanese quails and white leghorn chickens, incubation will happen in 'batches' (due to reduced egg production/supply). As a result, an identical grouping procedure within these species will be applied." (line number: 345-350).

20. "Prior to each day, birds will be food deprived at 4PM". Would not make sense that food deprivation was different in each species, being proportional to their body mass? For example, for a canary being food deprived since 4pm of the previous day should result in more hungry, and perhaps more motivation to feed and engage with the apparatus, that for a herring gull which has a larger body mass, as for a canary the costs of being food deprived will be larger, considering the same amount of time. Furthermore, please clarify whether birds will be allowed to feed ad libitum after the end of the habituation and testing periods until 4pm in each day. Finally, if in each day different individuals are tested, will the birds continue to be food deprived until the end of the habituation/testing periods? If not, wouldn't this influence the willing to feed as well, as they will feed more before the start of the next food deprivation period.

As mentioned in our reply to Comment 4 of Reviewer 1, we have adjusted the period so that it is more in line with previous work (for the canaries, see e.g., Müller et al., 2008) or with the standard procedures at the Wildlife Rescue Centre (where most of our research is taking place). We have also further clarified this in the main manuscript.

We would also like to stress here as well that we not interested in general differences between species per se, but rather in the interaction between species and barrier type. We deem it unlikely that such an interaction could be caused by (species) differences in willingness to feed.

Procedure

"[…] Food is provided *ad libitum*, but in the evening before an individual's habituation or testing day, the feeders will be removed from the enclosures at 6PM (after the last feeding time). This will create a non-feeding period during the night (which is normal and also happens in non-experimental conditions), followed by (shortly) delayed feeding in the morning to prevent birds from overindulging prior to habituation or testing. This is in line with other studies using the same species (chicken: e.g., Bollweg and Sparber, 1998; quail: e.g., Ueno and Suzuki, 2014 and unpublished data from our lab; herring gulls: e.g., Dewulf et al., 2022; domestic canaries: e.g., Müller et al., 2008). After all individuals of one enclosure have completed the habituation or testing trials for the day, food will be again provided *ad libitum*." (line number: 324-331).

21. "Sessions" are mentioned throughout the proposal, but only in the statistical analysis a clear indication of what different sessions would be exists: indicating that session 1 uses one type of bars in the barrier, and the session 2 the other type of bars. Before in the predictions section and Table 1, there is a slight indication that a session could be trials using the same barrier type, but not explicitly. But then in the video recording analysis, sessions are used to distinguish habituation vs 2 test sessions.

We agree that the term ‘sessions’ was inconsistently used in the manuscript. We have (explicitly) clarified this term as one or more trials using the same barrier type (without an interruption) on a single day. Thus, a trial is a single ‘attempt’ at the detour task, irrespectively whether the food is placed in front (habituation) or after (testing) the barrier whereas a session is one or more trials using the same barrier type on a single day.

Introduction

“ […] and the number of sessions per barrier type fluctuated between species (e.g., yellow-legged gulls received three sessions per barrier type spread over three days, while hybrid quails received one session per barrier type).” (line number: 123-125).

“In our partial replication, we will make several changes to address commonly raised concerns in the detour literature (including the concerns raised in the previous section, see table 1). […] All species will be given an equal amount of trials and sessions per barrier type (see below).” (line number: 141-146).

Predictions

”Second, as each session will consist of three trials (of the same barrier type), we can also look at how detour performance improves within each session. Based on previous studies, we predict that detour performance will improve across trials within a session (**Prediction 2**)” (line number: 192-194).

Procedure

”On the two testing days (10:30 AM - 02:30 PM), each bird will perform one session, each consisting of 3 trials with one barrier type. The order of barrier type (i.e., horizontal-bar or vertical-bar barrier) will be pseudo-randomized within and between species, across the two testing days.” (line number: 341-343).

Video Recording and Analysis

“The videos of the second and third habituation trial and the three test trials per test session will be coded using the free, open-source ’Behavioural Observation Research Interactive’ Software (BORIS, v.7.13.6) (Friard and Gamba, 2016). We will code five (types of) events (see table 3 and figure 3): latency to leave the start box (for the 2 habituation trials and the six test trials), persisting (test trials only), moment of detouring the barrier (test trials only), interacting with the food bowl (2 habituation trials and the six test trials) and leaving the species-specific 'test box zone of interest' (test trials only) […] ” (line number: 373-378).

22. I am not sure if I missed this in the statistical analysis description, but will the social group be controlled for? As I understood there must be 6 groups of 10 individuals each, for each species. I would imagine that group variation could be accounted for as random effect. If not, could the authors explain their decision?

Thank you for raising this issue. As noted above, we had (initially) opted for an AN(C)OVA as both Regolin et al., (1994) and Zucca et al., (2005) utilized a similar data analysis approach. An AN(C)OVA does not allow us to model random effects. However, we agree that (G)LMM’s are more flexible in assigning variance and estimating effects (including, adding a random effect of social group). Therefore, we have changed our data analysis plan (after further consultation with a statistical expert).

Statistical Analysis

"[…] Models will be fitted my means of the *lme4* package (Bates et al., 2015) and parameter estimation and p-values for the generated models will be provided by means of the *lmerTest* package (Kuznetsova et al., 2017) via the Satterthwaite's degrees of freedom method (linear mixed model, LMM) or via the *carData* (Fox, Weisberg, and Price, 2022) and *car* (Fox and Weisberg, 2019) packages (generalized linear mixed model, GLMM). For the (G)LMM, we will use partial eta-squared (η2p) as effect sizes and they will be calculated by means of the *effectsize* (Ben-Shachar et al., 2020) package.

A (G)LMM with Type III sum of squares will performed on the latency to detour and the cumulative time spent in the species-specific 'barrier zone of interest' (persisting). Both models will include the between-species factor: *Species* (i.e., white leghorn chickens, Japanese quails, herring gulls and canaries) and both within-species factors: *Barrier* (i.e., vertical- and horizontal-bar) and *Trial* (i.e., 1-3), and their interactions. In addition, we will add two extra explanatory variables to the model: a *'multi-baseline'* measure of an individual's general motivational state (and its interaction with Species, as we will mean-center this *'multi-baseline'* measure within species, see Chen et al., 2014 for an example of within-group centering); and *Barrier Order* (with two levels: did the individual receive the horizontal-bar barrier on the first test day 1 and the vertical-bar barrier on the second test day; or vice versa), as species might demonstrate superior performance with the last encountered barrier, irrespective of its type and ecological validity. Individual birds and enclosure (social group) will be included as a random intercept in the models, with individual birds nested in enclosures. In addition, we will include by-individual (nested in enclosures) random slopes that can vary for the levels of *Species* (corresponding with species-specific intercepts) (line number: 421-441).

On OSF, we have created a folder "AnalysisPlan" which contains a R-script with the to-be-used model (assumption) functions, and specific packages needed per function.

Note that we will now conduct post-hoc linear contrasts (instead of paired-t-tests, as was the case in the first draft of this Stage 1 RR), as our statistical advisor argued that performing post-hoc linear contrasts upon the model are more powerful (due to a more accurate estimation of the error).

23. In the last sentence of the predictions section sound a little bit too vague. Could the authors add a little on what they can predict on this? At least why having a significant three-way interaction could make sense or not in the context of this work.

We have clarified our explorative prediction three in the manuscript.

Predictions

"Furthermore, we will *explore* if the learning effect (i.e., improved detour performance across trials) interacts with the ecological validity of the stop signals. There are two possible patterns that would result in a three-way interaction between *Species*, *Barrier* (horizontal- vs. vertical-bar barriers), and *Trial* (1-3) (**Explorative Prediction 3**). First, detour performance might be better for ecologically valid compared with non-valid stop signals at the beginning, but this pattern might diminish over time as individuals learn

to stop (i.e., the difference between barrier types would decrease). Second, detour performance might be poor at the beginning for both barrier types, but learning to stop might be easier for ecologically valid signals compared with non-valid stop signals (i.e., the differences between barrier types would increase). Both patterns would be theoretically meaningful, but we do not have *a-priori* predictions about the direction of the three-way interaction." (line number: 196-204).

24. One last thought, do the authors expect any influence of the fact that all birds in each species did not experience a wild or more natural environment prior the experiments. Furthermore, do the authors expect an influence from differences in development social context prior being included in the groups of 10 individuals.

The previous studies (which we aim to replicate here) argued that detour performance is improved when the perceptual characteristics of the barrier (the stop-signal) match the original **ecological niche** of a bird species (*Species* x *Barrier* interaction effect). As the authors of the original studies referred to the ancestral environments of the species, we assume that these effects should also be present without experiencing a wild or more natural environment. Note that in the original studies, species also did not experience a more natural or wild environment prior to the experiments. Of course, it is possible that experience with a wild or more natural environment would further strengthen the *Species* x *Barrier* interaction effect, but such post-natal effects are not the focus of the present study.

With regards to the social context: this might contribute to some general differences between species, but we deem it unlikely that social context could account for the critical interaction between species and barrier type.

<u>Minor comments:</u>

25. I struggled a bit to understand what is specifically being stated in the first sentences of the 3rd paragraph of introduction.

We have clarified the first sentences of the 3rd paragraph of the introduction in the manuscript.

"[…]. Typically, performance in the detour task has been linked to the variation in the effectiveness of a single cognitive control function, 'response inhibition', or more generally, 'inhibitory control' (which is an umbrella term for various types of inhibition, which may or may not be related to each other; Bari and Robbins, 2013). However, by referring to general, ill-defined cognitive constructs such as RI (or even worse, a general umbrella term such as 'inhibitory control'), we do not explain the underlying cognitive mechanisms or building blocks of stopping (Verbruggen, McLaren, et al., 2014), as the explanation is 'just as mysterious as the thing it is supposed to explain' (Press et al., 2022)." (line number: 40-46).

26. in "sample size" section, why are 60 individuals "the largest number that is practically feasible"? Is this only according to the authors' aviaries conditions constraints?

Yes. We have clarified this in the manuscript <u>and</u> study design template, column "Sampling plan".

Sample size

"We will test 60 individuals per species. A-priori power sensitivity analyses done in G*Power (Faul et al., 2009) indicate that this is sufficient to detect small effects; it is also the largest number that is practically feasible \footnote{Farrar et al. (2020) mention in their paper on replications in comparative cognition that power analyses are not the golden standard in this research domain, and 'in many cases comparative cognition researchers could be better off performing design or sensitivity analyses based on their resource constraints."}) […]" (line number: 213-215).

Study design template

"We will test 60 individuals per species (total N = 240). *A-priori* power sensitivity analyses done in G*Power (Faul et al., 2009) indicate that this is sufficient to detect small effects; it is also the largest number that is practically feasible given our resource constraints and study design (see Farrar et al., (2020)) […]"

27. I think some explanation should be given on how the maximum testing times were defined, and why the times are different between habituation and testing periods.

We have clarified the maximum habituation and testing (and why they differ) in the manuscript.

Procedure

"[…] Maximum trial times during habituation will be longer than during testing, as the main goal of the habituation is to familiarize each bird with the test material (and obtain a 'multi-baseline' measure of an individual's general motivational state). The maximum duration of a test trial will be 2 minutes (after an additional 15 seconds inside the start box with the second, transparent door), which is in line with other studies (e.g., Vernouillet at al. 2016; and Kabadayi, Krasheninnikova, et al. 2017). Two minutes should be sufficient, especially because our barriers are not entirely transparent (hence, will partially occlude the food reward), making it easier to execute a detour response (Kabadayi, Bobrowicz, et al., 2018)." (line number: 360-366).