Dear PCI RR Managing Board,

Dear Recommender Dieter Lukas,

Dear Editor,

We would like to thank you for the opportunity to revise and resubmit our Stage 2 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 constructive feedback. In this Stage 2 resubmission, we have carefully addressed all comments and suggestions in our responses below, where line numbers correspond to the manuscript with tracked changes. Updates to the Methods section (Statistical Analysis subsection) since Stage 1 IPA are highlighted in orange to clarify deviations from the original peer-reviewed analysis plan and prevent any potential confusion about aligning the data with our purposes (see comment 1). For all other sections, changes made since the Stage 2 – Version 1 (Stage 2.V1) report are highlighted in red.

The most substantial change concerns the Method section. We thoroughly revised this section to further clarify and motivate deviations from the original Stage 1 peer-reviewed analysis plan (comment 1), and addressed and corrected an inappropriate calculation and visualization of residual autocorrelation (comment 3). We also resolved ambiguities in the manuscript, including those related to minimum trial durations, the temporal correlation structure of the model for detour latency, the sample of birds across trials, and the specifications of the food bowl. Finally, we refined key elements of the Discussion section, such as the importance of examining both dependent measures and the broader relevance of our study.

We would like to disclose that we noticed the absence of a seed in our R scripts (required for reproducibility). This led to slight changes in parameter estimates ($< \pm 0.01$, except for the intercept) for the ZINB model (i.e., the model for persisting) when rerunning the scripts, due to the stochastic components involved in the parameter optimization process. To address this, we included a seed for reproducibility, reran all analyses, and updated all relevant R scripts (*Analysis_Detour_LatencyV2.R*, *Analysis_Persisting_V2.R*, and *Analysis_GroupSize_ExtraV2.R*), which are now available on OSF. Changes in parameter estimates are highlighted in red within the tables in the manuscript. The other models were not influenced.

We declare that this revised Stage 2 RR remains original and unpublished. All authors have approved the submission of the revised Stage 2 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 2 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

1. [...] However, I was confused about the changes to the statistical approach. On first reading, my impression was that you had made changes that were not preregistered. After checking the supplementary material, I now think these changes were mostly covered by the initial plans to change the statistical approach in case of "potential violations of model assumptions" (text included in the Stage 1 report). This does need to be clarified though in the manuscript. Currently, reading the methods of the Stage 2 manuscript can give the impression that you decided to change your analyses after you had seen the data to best fit your purposes. This would represent a major deviation. There are three potential issues here of having such a change appear in the Stage 2 manuscript. First, if these are additional analyses that were not peer reviewed, they would need to appear as exploratory analyses in the discussion, rather than as the main results. Second, it could affect the level of bias as originally this report had a level 6, but that would clearly be lower if decisions were made based on observing the data. Third, it is unclear whether you tried additional analyses that you did not report.

To avoid these issues, I therefore ask you to revise the method and result sections of your manuscript to make it clear how these deviations were developed, and how they fit within the original plan that was peer reviewed. Some of this information is in the supplementary material, but it is not always referenced and I am not sure whether all the information is included for how a particular outcome lead to a particular decision. I think it would be necessary to report in the manuscript the assessment of the primary analyses you performed to show that they are not reliable. In the methods, you should also preface each of the changes you did with whether it occurred according to the potential violations you mentioned in the Stage 1 report.

The Method section (Statistical Analysis subsection) of the manuscript has been revised. For reviewing purposes, we highlighted changes and modifications compared to Stage 1 IPA (not Stage 2.V1) in orange. We have further described and justified these changes to remove any potential confusion about fitting the data to our purposes. Additionally, we have updated the titles of the *Results* section to more accurately distinguish between registered comparisons (main analyses) and additional exploratory analyses.

The evolution of the model structure (from the registered to the applied model) and visualizations of model assumption violations for all models are provided in the supplementary materials. Furthermore, the data-analysis pipeline, including the assessment of primary analyses, model assumption checks, and any resulting adjustments, is thoroughly documented in the R scripts *Analysis_Detour_Latency.R* and *Analysis_Persisting.R*, which are available on OSF.

To further enhance reproducibility, we have also uploaded two updated scripts to OSF: *Analysis_Detour_Latency_V2.R* and *Analysis_Persisting_V2.R*, with changes from the Stage 2.V1 submission clearly marked. These scripts facilitate the replication of the diagnostic plots included in the supplementary materials (e.g., to demonstrate violations of model assumptions). While the plots are very similar to those originally used for model selection,

they have been refined for inclusion in the manuscript and can be reproduced as presented, provided the data have been sorted accordingly (see comment 3).

Due to the length of the revisions and the complexity of demonstrating changes between the Stage 1 IPA and subsequent modifications, we have not included the revised text in this letter. However, the revised sections can be found on lines 440–538.

2. There are two changes you did to the statistical approach where I have more specific questions. The first is your following decision: "A minimum trial duration of zero seconds for persisting was assigned to the 483 trials (33.82% of the data set) in which birds did not enter the species-specific 'barrier zone of interest'." I assume these are birds who nevertheless detoured and interacted with the food bowl? Or does this include individuals who received a zero value because they never left the area around the start box?

We would like to clarify that birds that neither detoured nor entered the species-specific "barrier zone of interest' during a test trial (e.g., birds that never left the area around the start box), were excluded from subsequent test trials, and their data were not included in the statistical analyses, following the mid-test exclusion criterion 1. For further details, please refer to lines 418-420 in the manuscript.

We agree that the way of reporting the sample of birds that received the minimum and maximum trial durations for persistence and detour latency respectively was ambiguous, and we have clarified this in the manuscript:

> Video Recording and Analysis
> "A maximum trial duration of 135 seconds for detour latency was assigned to the 20 trials (1.40\% of the data set) in which birds did not to detour but entered the species-specific 'barrier zone of interest'." (line number: 401-403).
>
> "A minimum trial duration of 0 seconds for persisting was assigned to the 483 trials (33.82\% of the data set) in which birds detoured without entering the species-specific 'barrier zone of interest' first." (line number: 406-408).

3a. The second is about including a temporal correlation structure in the model of detour latency. It is not clear how this affects your interpretation, because as far as I understand it your measure of time is the trial? Accordingly, you have trial in the model repeatedly? Does this temporal correlation structure reflect individual level differences?

The correlation structure accounts for autocorrelation in repeated measures over Time (i.e., trials 1-6; for each bird, nested within enclosures). This structure does not directly account for individual differences, as a random slope would, but rather captures the temporal dependencies in the data. Specifically, each bird participated in two sessions, with one session per barrier type and three trials per session, resulting in six interdependent trials. The structure assumes a decay of correlation over time, meaning that the correlation weakens as the time interval between trials increases.

Thus, this correlation structure models the dependency of the residuals that are close to each other with respect to the time of measurement. We included this because neither the main effect of Trial (i.e., trials 1-3 within a session) nor the Barrier (Horizontal-barred vs. Vertical-barred; pseudo-randomized] × Trial interaction can adequately account for these dependencies over all 6 trials.

We admit that the calculation and visualization of the autocorrelation structure, shown in Stage 2.V1, may have been misleading and potentially inappropriate for interpreting our data and models for two reasons.

First, after further consultation with a statistical expert, we came to the conclusion that the Stage 1 registered Durbin-Watson statistic and the Stage 2.V1 ACF plot (using the base R package) are actually not suitable for (G)LMM due to the random effect structure. In the previous manuscript, the ACF plot was incorrectly calculated for time series data, with correlation between residuals assessed beyond 5 lags. This was problematic because temporal autocorrelation should only be examined within individuals over time, not between birds. Given that birds are independent and each had only 6 trials, temporal autocorrelation should only be assessed up to lag 5.

To address this, we used ACF plots with standardized residuals across the six time points (Time or total number of trials) for each bird (for the LMM modeling detour latency) and simulation-based residual plots to inspect autocorrelation between residuals (for the GLMM modeling persistence).

Second, in Stage 2.V1, we overlooked the need to manually sort the dataset by bird and time point (we incorrectly assumed that the used functions would sort the data within birds automatically). We have addressed this issue as well in the revision. Importantly, this sorting issue does not impact the estimates of the parameters, the significance levels, or the overall model fit.

In the revised manuscript, we have reworded the section on our applied detour latency model to make it easier to understand the need to include this temporal correlation structure. In addition, a footnote has been added to address the inappropriate autocorrelation checks in the Stage 1 and Stage 2 V1 manuscripts and to outline the alternative methods implemented.

> Applied model
> For detour latency, [...]To address autocorrelation in the residuals \footnote{The Stage 1 registered Durbin-Watson statistic (using *performance* Lüdecke et al., 2021 package) and the ACF plot (using the base R package) are not suitable for (G)LMM (due to the random effect structure) an issue identified during the data analysis. Consequently, we opted for alternative methods to assess autocorrelation in the residuals. Specifically, we employed ACF plots (for LMM; using the package *nlme*, Pinheiro and DM Bates, 2000) or simulation-based residual plots (GLMM; using the package *DHARMa*, Hartig, 2022) for inspecting autocorrelation between residuals.} the model was further extended with a temporal correlation structure using the *nlme* package (Pinheiro and DM Bates, 2000). This temporal correlation structure accounts for the correlation in residuals from repeated measurements across Time (i.e., 1-6 trials; for each bird, nested within enclosures). Specifically, each bird participated in two sessions, with one session per barrier type and three trials per session, resulting in six interdependent trials. The autocorrelation parameter (ϕ), estimated by the model at lag 1, was 0.319. Explicitly modeling this autocorrelation properly accounts for the residuals' temporal dependencies (see supplementary Figure 2, Dewulf, Garcia-Co, et al., 2023) leading to an improved model fit (AIC = 4063.716 with the correlation structure vs. AIC 4275.634 without) and more accurate parameter estimates. (line number: 488-496)

We have also updated the residual diagnostic plots in the supplementary materials accordingly. Two scripts, titled *'Analysis_Detour_Latency_V2.R*' and *'Analysis_Persisting_V2.R*,' are available on OSF, allowing the plots to be replicated within the data-analysis pipeline.

3b. I was also confused because the supplementary materials appear to indicate that this correlation is negative - so when an individual performed above average in one trial, it performed below average in the next trial but at average several trials later? Why would there be a negative correlation, if anything I would expect a positive one?

After redoing the analyses, we actually observed a small positive phi value (modelled autocorrelation parameter) of 0.319 (calculated on lag 1). This small positive correlation indicates that if a bird's observation exceeds the model's expectation in one trial, it is likely to do so again in the next trial; similarly, if the observation is below the model's expectation in one trial, it is likely to be below in the next.

Concerning the remaining autocorrelation that was not explicitly modelled, the ACF plot shows autocorrelation values ranging between -0.2 and 0.2 (except for lag 5, but which was based on a single comparison between residuals 1 and 6 and is therefore less reliable). Given their small magnitude, they can be considered negligible and do not warrant further model adjustment.

Thus, we can conclude that there is a small positive correlation between consecutive time points (or trials, as measured on lag 1). Given the relatively low, unmodeled autocorrelation values, as visualized in the ACF plot, there is no need to explore more complex models that account for correlations between further lags.

4. I have one more question about the interpretation: your decision to exclude birds according to the different criteria means that your sample during Trial 1 is not the same as your sample during the final Trials. Do you think this could lead to selection biases that affect your results? In particular, could the increase in performance be slightly explained by birds who take longer initially being more likely to drop out? For example, if there are four birds of whom three take 5 seconds and one takes 10 second during the first trial, but the slowest bird later no longer participates, the average from Trial 1 to Trial 2 would drop from 6.25 seconds to 5 seconds. As far as I understand, you include an individual level offset for detour latency, but not individual level slopes of the change across the trials?

We would like to clarify that the sample of birds remains consistent across trials. Our exclusion criteria ensure that birds excluded from further testing are also removed from the statistical analyses. For further details, please refer to lines 418-420 & 426-428 in the manuscript.

We would also like to clarify that we included an individual-level offset for detour latency, represented by the random intercept term (1 | ID:group). We agree that individual-level slopes for the change across trials can still be included, even with consistent samples across trials. To explore this further, we ran an additional model that incorporated both the individual-level offset for detour latency and an individual-level slope for trial-to-trial changes, specified with the random effect structure (1 + Trial | ID:Group). However, this extended model (AIC 4083.322) did not outperform the reported model in the main manuscript revision (AIC:

4063.716) for detour latency. The R script "*Stage2V1Comment4.R*" contains the extended models and output and can be used to replicate them.

Comments Christian Nawroth

5. L618-619: Please explain why this demonstrates the value of analyzing two parameters.

We have clarified this in the manuscript.

> Discussion
> "However, this pattern demonstrates the value of looking at detour latency and time spent interacting with the barrier. One might assume that lower persistence scores should automatically result in shorter detour latencies but for gulls, this was not the case. This indicates that overall task performance (i.e., detour latency) captures additional behaviours, potentially unrelated to response inhibition (e.g., the time taken to approach the barrier, time spent not interacting, time needed to navigate the barrier, etc.)." (line number: 668-673).

6. L452-453: There is a missing closing bracket in "(corresponding with species-specific intercepts."

We have included the closing bracket in the manuscript.

7. L497: For consistency, "Barrier x Species x Trial" should read "Species x Barrier x Trial."

We have updated the manuscript, including all predictions, R output, and supplementary materials, to ensure that all interaction terms—whether two-way (e.g., Species x Barrier, Species x Trial, Barrier x Trial) or three-way (e.g., Species x Barrier x Trial)—are consistently presented with Species listed first, followed by Barrier, and then Trial. These changes have not been highlighted in red as they simply reflect a reordering of existing terms for consistency.

8. L611: The phrase "inhibit their initial behavior" is somewhat vague. Could you clarify what is meant by "initial behavior" in this context?

We have added the clarification to the manuscript.

> Discussion
> "[...] individuals had to learn both to inhibit their prepotent response to go directly for the reward (as the direct path is blocked) and to navigate around the barrier [...]" (line number: 660-661).

9. L622: There should be no comma after "latency."

We have removed the comma in the manuscript.

10. L622: Could you specify what kind of subcomponents are being referred to?

We have clarified this in the manuscript (but see also our reply to comment 5)

> Discussion
> "First, the fact that gulls showed evidence of learning in measures of persistence but not in the measure of detour latency suggest that, at least for some species, certain task components are more influenced by learning (inhibiting an unrewarded repetitive response) than others (inhibiting the response to go straight for the food or navigating around a barrier, which are both captured by the detour latency)." (line number: 673-678).

11. L624: It would be helpful to include a brief example of how this might relate to the species' ecological niche.

We have included a brief example in the manuscript.

> Discussion
> "Speculatively, this could be related to the ecological niche adaptations of the species as well. Certain behaviours, such as inhibition of unrewarded responses, may be more critical than others in certain ecological niches, making them easier to learn. In contrast, other behaviours, such as navigating obstacles, may be more influenced by context-specific factors, and therefore, harder to learn for certain species (although follow-up work is required to test this idea)." (line number: 678-682).

12. L654: A line break is needed here.

We have inserted a line break.

13. L662: Could you elaborate on how the exclusion criteria were specifically linked to the performance of the canaries?

We have included a speculative discussion in the manuscript on the relationship between exclusion criteria and the performance of canaries.

> Discussion
> "We speculate that canaries were able to solve the detour problem in our study, but not in the original work, due to the exclusion criteria we implemented, which ensured proficiency with the basic task demands (e.g., the perceptual, motoric, and motivational requirements for retrieving a food reward; Maclean et al., 2014) Specifically, our pre-test exclusion criterion ensured that all included birds visited and ate from a food bowl placed in front of a barrier (novel object) in the habituation phase before access to the food bowl was restricted by moving the barrier in front of it in the test phase. We believe that experience with retrieving the reward may be critical for measuring detour performance, potentially more so in aerially adapted birds. After all, Zucca et al. (2005) found that even after prolonged exposure to the test situation, a

large proportion of canaries were unable to solve the detour problem. This suggests that the problem was not a lack of familiarity with the test itself, but rather a lack of experience with retrieving the reward. However, this explanation is speculative and requires further investigation." (line number: 719-730).

14. L667: I fully agree with this statement. Perhaps you could additionally emphasize how this study particularly highlights its relevance for animal behaviour research.

We have added a new section in the manuscript that emphasizes the relevance of this study for animal behaviour research.

Discussion

"Although our study did not provide strong evidence for the idea that interspecies differences in the perception of barrier types influence detour performance (and cause species differences), this does not negate the need for further research into the influence of the characteristics of the stop signal or other underlying mechanisms of RI. More generally, future research should focus on the cognitive mechanisms underlying RI. Understanding these mechanisms will help explain inter-individual variation such as in decision-making in dynamic environments (Johnson-Ulrich and Holekamp, 2020), predator avoidance and foraging optimization (Tvardíková and Fuchs, 2012), as well as responses to broader ecological pressures (Lee and Thornton, 2021)." (line number: 734-741).

15. Table 4: The text notes, "The first 60 individuals (58 for quails) that did not fail any exclusion criteria were selected for this study, ensuring a balanced design and minimizing group variation." Could you provide a rationale for the substantially higher number of canaries included in the study (maybe as a footnote or similar)?

We have rephrased the note for Table 4. We included a larger number of birds in the experiment to ensure a consistent prior experience for birds participating in subsequent studies.

Table 4: Text note

All raised birds were subjected to habituation and (part of) testing. As can be seen, the total number of birds tested was higher than registered for all species (apart from the quails). This was due to the fact that these individuals were also reused for other studies, with different sample size requirements. Reusing individuals in other behavioral studies is possible when they share similar prior experiences (Van Horik, Langley, et al., 2018), and facilitates future analyses, such as exploring correlations between different tasks and making comparisons across studies. The first 60 individuals (58 for quails) that did not fail any exclusion criteria were selected for this study, ensuring a balanced design and minimizing group variation." (line number: 438).

16. Table 10 and beyond: Although the number of frames is mentioned in the text, it may be helpful to explicitly include the unit of measurement in this and other tables.

In the revised manuscript, we have included the appropriate scales and measurement units in each table heading of the descriptive statistic tables for both persisting and detour latency (for consistency).

> An example of the new heading of one of the descriptive statistic tables for detour latency (but see also table 7-8)
> The model predicted means (on the log scale), the back-transformed model-predicted means (on the original scale, in seconds) and the observed means (also on the original scale, in seconds) for detour latency across different Trial levels. (Table 6; line number: 551).

> An example of the new heading of one of the descriptive statistic tables for persisting (but see also table 11-16)
> Table 10. The model predicted means (on the log scale), the back-transformed model-predicted means (on the original scale, in frames) and the observed means (also on the original scale, in frames) for persisting across different Trial levels (Table 10; line number: 576).

17. Table 17: To improve comprehension, consider adding a note to clarify how, for example, only 3 quails came from a group of 6 individuals (e.g., accounting for dropouts).

For clarity, we have renamed Table 17 in our manuscript.

> Title Table 17:
> Visualization of the number of individuals that met our exclusion criteria, in relation to the enclosure group size and the species. (Table 17; line number: 632).

Comments Reviewer 1

18. Lines 61-62: is this evidence in humans, or animals in general?

The evidence we referred to is based on human studies. We rephrased this in the manuscript to avoid further confusion.

> The Crucial Role of Stop-Signal Detection
> "Several lines of evidence indicate that signal detection may play a critical role in RI, particularly in humans and non-human primates." (line number: 60-61)

Behavioural evidence in non-human animals, particularly birds, is discussed further in the following section. For further details, please refer to lines 69-83 in the manuscript

19. Line 261: how long is the initial indoor period?

The initial indoor period refers to the time spent in the heated rooms, where the chicks were housed in boxes with netting bottoms and hand fed for 5 days (and till their body mass exceeded 60 grams).

We have clarified in the manuscript that this refers to the same period, avoiding any further confusion.

Herring gulls
"Once hatched, the semi-precocial gull chicks [...] placed, in boxes with netting bottoms [...]) within heated rooms [...]. After this initial period in the boxes [...]" (line number: 256-262).

20. Lines 327-328: food was given in a "coloured food bowl". What colours were in this bowl? Could the different colour perceptions or sensitivities of the different species influence the level to which they reacted to the bowl?

We have added this information in the revised version, and explicitly added that the coloured food bowl during the 10-day habituation period in the enclosure was identical to the one used in the test box.

We believe that variations in color sensitivities or perception did not affect our results, as the extensive habituation period likely mitigated any initial differences and allowed the animals to learn the bowl-food reward associations. We have included the following explanation as a footnote in the manuscript.

Apparatus:
"[...] placed in a coloured bowl. For chickens and quails, these were coloured green and yellow (brand: Junai, The Netherlands); for gulls and canaries, these were coloured orange-brown (brand Elho, Belgium). \footnote{Potential variations in colour perceptions and sensitivities across species are mitigated by the developed preference for their respective food bowls prior to the start of the experiment through repeated exposure (i.e., 10-day habituation to the food bowl in the enclosure) and learning (pairing of the coloured bowl with food during these 10 days). }"
(line number: 305-307)

Procedure:
Prior to the start of the experiment, birds were habituated for 10 days in their enclosure to feed from a coloured food bowl, which was identical to the bowl used during both habituation and testing in the test box. (line number: 330-331)

21. Lines 347-348: I had this comment before and I do not think this is clarified yet. I understand clearly now that each individual performed only 3 trials overall (please correct me if I am missing something). Can you provide the reasoning of why only 3 trials were performed and whether this is representative to compare detour performance between species, when previous within species and comparative studies with detour tasks in birds considered a higher number of trials?

Each bird received two sessions (1 session/day) each consisting of 3 trials with one barrier type. We apologize for any confusion caused by the phrasing in the manuscript and have clarified this part in the revised version.

Procedure:
Each bird participated in one session per day on the two testing days (10:30AM - 02:30 PM). Each session consisted of 3 trials with one barrier type. The order of barrier type (i.e., horizontal-bar or vertical-bar barrier) was pseudo-randomized within and between species, across the two testing days. (line number: 351-354)

Our Stage 1 decision to include only 3 trials per session was based on previous work indicating that most learning typically occurs within the first few trials (e.g., Logan, 1988; Thorndike, 1913; see e.g. Van Horik, Langley, et al., 2018, for a detour example). As this study included four species and a relatively high number of individuals per species, we therefore decided not to include additional trials for reasons of feasibility. Furthermore, additional trials would have complicated the interpretation of the results by increasing the likelihood of order effects (e.g., horizontal vs. vertical bar barriers presented first).

22. Lines 249-253: would it make sense to include lifespan as a covariate in models to control for this possible developmental effect?

Adding lifespan as a covariate in our model would introduce multicollinearity with species, as each species is associated with a distinct lifespan (e.g., gulls: 49 years, canaries: 24 years, chickens: 17.5 years [average], quails: 6 years). Therefore, we did not do this.