Comment 1:

Jeremy Van Cleve

The authors use wild caught great-tailed grackles to study how behavioral flexibility is connected to behavioral traits related to boldness, exploration, motor diversity, and persistence. The authors restrict their analysis to traits that are repeatable and find that exploration and persistence were repeatable and both of these had some relationship with flexibility. The study has a registered and peer-reviewed preregistration manuscript and the authors detail how their analyses matched or deviated from the preregistration.

Two reviewers have assessed the manuscript and both have a number of suggestions for improvement. The biggest issue the reviewers note is how the manuscript connects to the preregistration manuscript. Both reviewers have some difficulty reading the manuscript without explicitly referring back to the preregistration. In so far as the authors would like the manuscript to readable on its own without necessarily reading the preregistration first, they should consider the reviewers comments on this issue. I agree that the overall readability was impaired a bit from referring back to the preregistration manuscript and describing differences with the preregistration. I think the readability would be improved by having the main body describe the methods and results in whole and then leaving differences with the preregistration to supplementary material. One reviewer mentions dropping the wild vs captivity comparison; this may help with the readability though the authors may find a place for these results in the revision.

The other suggestions by the reviewers are mostly related to adding clarification about the conceptual relationships between the behavioral traits and clarifying some points in the analyses and discussion. A revision can easily accommodate or address these suggestions.

Response 1: Thank you for the feedback and taking time to review our manuscript and coordinate with the reviewers. We hope you find the revision much improved in readability.

Comment 2:

Anonymous reviewer 1

In this paper, the authors test whether and how measures of cognitive flexibility relate to other (repeatable) behavioral traits such as those typically interpreted as “exploration” or “boldness” using grackles. To do this, they first tested for repeatability in these other behavioral traits and found that no measures of “boldness” (approach to putatively threatening objects) were repeatable. However, they did find that some behavioral measures in the exploration assay were repeatable: birds were highly repeatable in their latency to approach and spent time near a novel environment, but not novel objects. They also found that the birds were modestly repeatable in the number of touches they made to all the different objects across the assays. Then once the authors established that certain behaviors were repeatable, they looked for relationships between those behaviors and specifically ‘cognitive flexibility’. They found no

relationship between flexibility and the exploration measures. They then also looked for relationships between persistence and flexibility and found that more flexible birds were more persistent in some ways, but not other ways. Altogether, I actually had a bit of a hard time reading the manuscript. I really genuinely appreciate the fact that authors are clearly making a big effort to use research best practices in pre-registering their methods and questions. However, I don't know if this is what is expected with the post-study write-ups, but this manuscript is pretty difficult to evaluate as a stand-alone document. I'm not sure if the general expectation is that any reader would first read the pre-registration and then read the post-study write-up (i.e. manuscript) but that seems to be necessary here in order to evaluate this manuscript as many of the important methodological details are missing. *So my biggest comment is that I think the authors should really put the relevant parts of the methods that were included in the pre-registration into the manuscript here so that the work can be evaluated in its entirety with just one document. The authors can describe their methods as planned and then explain when/how they deviated from these plans as necessary.*

Response 2: We appreciate you taking the time to review our post-study manuscript. Ideally, the reviewers of the post-study are the same as those for the preregistration, and therefore have some familiarity with the preregistered research plan. The reviewers then assess the post-study manuscript for whether the authors conducted the study in the way that it was preregistered. For this study, there was a long delay between the preregistration and post-study manuscript, so even if the reviewers are the same, they might not remember the preregistration. Therefore, we value your comments about where the post-study manuscript requires more clarity.

To address this main concern, we added significant details to our methods. These additions are explained in Responses 4, 21, & 22 below.

Comment 3:
Just wanted to mention that I really appreciate the motivation of the study – to disentangle the jingle-jangle fallacy with these behavioral measures. This is something I think about a lot – we interpret different behaviors (e.g. latency to approach an object) as different/similar things ("exploration" versus "boldness") which makes major assumptions about what we think are the underlying causes of variation in those behaviors. I really would love to see more work that comprehensively tries to figure out what the actual axes of behavioral variation are as opposed to a priori assuming we know them!

Response 3: Thank you so much for the positive feedback!

Comment 4:
At the start of the methods it would be super helpful to have some sort of overview – how many birds were used in total, how many times were they measured, in what behaviors? Reading the MS, I think that 19 birds were used total, and they were measured twice in the 'boldness' and 'exploration' assays but I'm struggling to figure out how/when motor diversity/persistence was

measured and how many times cognitive flexibility was measured? Were these assays performed on the same days? Different days? How long were the birds in captivity?

Response 4: We added more details to remedy this confusion in the first two paragraphs of the methods ("Subjects" and a new paragraph with the heading "Test battery"), Lines 169-199.

Comment 5:
Line 281: For the repeatability analyses, what other fixed effects were included in the models? The authors report the (I think, adjusted) repeatability estimates so having the authors also report the marginal and conditional R-squared values from their models would be super helpful. E.g. a repeatability of 0.85 seems high, but if the fixed effects of the models are actually explaining 80% of the variance in the data, then this means that individual differences only accounts for 85% of the remaining 20% of the data.

Response 5: Yes, we presented adjusted repeatability because there were several covariates that we needed to control for. We clarified this in Lines 294-296 (changes in bold):

"For each behavioral trait, we included fixed effects to control for variation in the response not attributable to individual differences **and consequently we report the adjusted repeatability estimates.**"

And also Lines 300-302:

"Marginal and conditional R-squared values are reported in Table S1 of Supplementary Material 2 to illustrate the impact of fixed effects on repeatability estimates."

In our models for the two dependent variables representing exploration of the novel environment, we included a covariate describing whether the grackle was flexibility trained or in the control group (FlexGroup). Based on Nakagawa & Schielzeth 2013, we checked the marginal and conditional R-squared values in the repeatability models. We found this covariate accounts for a small portion of the variance in Latency to approach (R2 marginal = 0.05, R2 conditional = 0.71), or Duration near the novel environment (R2 marginal = 0.19, R2 conditional = 0.89). Similarly we only had one fixed effect (FlexGroup) in our model for exploration of the novel object that accounted for much of the variance (R2 marginal = 0.10, R2 conditional = 0.11).

In the boldness models, there was a covariate for FlexGroup and one for boldness condition (hawk or cat). These covariates did account for some variance with the R2 marginal = 0.07, and the R2 conditional = 0.14.

In the persistence models, we included covariates for FlexGroup, assay type, and total time grackles spent with each assay to account for the opportunity to make touches. The covariates did account for much of the variance, R2 marginal = 0.57 and R2 conditional = 0.69.

Lastly, the motor diversity repeatability models included covariates for FlexGroup and assay type. The covariates accounted for much of the variance with the R2 marginal mean = 0.06 and the R2 conditional mean = 0.08.

We added these values to Table S1 in the Supplementary Material 2

Comment 6:
Line 290: How exactly did you test if these behaviors were 'correlated' with flexibility? Did you just run a simple pearsons (or spearmans) correlation? Or did you use a bivariate model to account for the repeated measures on the individuals (this would be the most appropriate thing to do).

I recognize that much/all of this information may be in the pre-reg but it really also needs to be included here.

Response 6: We assessed correlation by interpreting the parameter estimates from Bayesian multivariate models. Specifically, these models included exploration or persistence as the dependent variable and, as independent variables, the Flexibility Comprehensive variables OR whether the grackle was in the control or trained group for serial reversal learning. We also included a random effect for bird ID to account for repeated measures, and for the persistence model we included an offset term for time to account for different trial durations leading to variation in opportunity to contact the multiaccess boxes. To clarify this in the text, we updated the section Methods > Statistical analyses > Relationship with flexibility (additions indicated here in bold face font), Lines 310-317:

"If performance was repeatable across two time points in the behavioral trait assays, **we used the average value per bird per assay in Bayesian multivariate models** to investigate whether performance was **related to** the Flexibility Comprehensive variables (ϕ and λ). **As such, the performance variables from each behavioral trait assay were the dependent variables and ϕ and λ were the independent variables. We assessed the relationship between flexibility and the behavioral trait by interpreting the parameter estimates from these models. Similarly, we used Bayesian bivariate models** to analyze whether there was a difference in performance on the behavioral trait assays between grackles that underwent serial reversal learning flexibility training relative to grackles in the control group."

Comment 7:
I recognize that the authors pre-registered this so their hypotheses/predictions already went through review, but I just wanted to note that the hypotheses they listed in the manuscript are actually not really (biological) hypotheses. Biological hypotheses are explanations for why something is the way it is. So for example, hypothesis 1 simply states that behavioral flexibility will be correlated with exploration but not boldness, but doesn't say why. The why bit is the

hypothesis and would be the interesting bit to understand! The authors do have hints of hypotheses in some of their alternative predictions (e.g. P1 alt 4 states "[no correlation between exploration and flexibility may happen because]…these measures of exploration incorporate novelty and thus measure boldness rather than exploration." This bit here is an EXPLANATION for why there may or may not be a relationship between these measures. The second halves of P6 alt 1, 2, 3 also contain potential hypotheses. So really the things that authors have written as hypotheses here are in fact just statistical null hypotheses which are not biologically that interesting (e.g. H0: there is no relationships between X and Y; H1: there is a relationship between X and Y) but are not biological explanations for why that relationship is happening. I just mention this because I see this as a good opportunity for the authors to make their paper stronger and more impactful!

Response 7: We understand that a popular and well-supported form of a hypothesis includes a biological explanation. However, one of the main goals of preregistrations is to increase replicability of the methods and so the hypothesis/prediction structure focuses on the testability of the question through specifying relationships among variables (more information on this process is given at PCI Registered Reports, in the description of the study design table here). Including explanations of the biological phenomenon governing these relationships can instead be detailed in the discussion because we did not specifically measure those biological phenomenon in this experiment (e.g., we did not measure in this manuscript whether flexibility and exploration affect adaptation to human-induced environmental change).

You are correct that this preregistration was peer-reviewed and approved with the current hypothesis/predictions structure and so we can't change our hypotheses and predictions for this Stage 2 version of the preregistered manuscript. Instead we emphasized in the discussion the biological interpretation of our results supporting our hypothesis in Lines 555: "in support of our main hypothesis, we did find a relationship between behavioral flexibility and exploration",

and 602-605: "This potentially explains how great-tailed grackles are successful at adapting to rapid anthropogenic change. The individuals in the population that are willing to seek out novel foraging or nesting opportunities are also able to change their behavior to switch to using these novel resources when they are encountered."

Comment 8:
This is something the authors can't control and I'm sure they're aware of it. But they are asking a lot of 19 data points really… For the number and complexity of analyses they are doing, it seems unlikely they have the power to detect anything but the strongest effects.

Response 8: The power analysis from our preregistration predicting the ability to detect a relationship between behavioral trait measures and flexibility shows that with a sample size of 32 we have a 70% chance of detecting a large effect. We did not end up with a sample size of 32, but we also have significantly fewer predictor variables (2 instead of 10). When we re-ran

the power analysis with our observed sample size and predictors, we have a 54% chance of detecting a large effect. One motivation for doing the flexibility training through serial reversal learning was to increase the effect size for the relationship between flexibility and other behavioral traits, if it exists. Nevertheless, we recognize this smaller sample size is not ideal and highlight this in the discussion in Lines 616-626:

“In addition, with a sample size of 19, we potentially lacked the power to detect a subtle relationship between flexibility and exploration or persistence. We conducted a power analysis *a priori* that indicated that a sample size of 32 would permit detections of large effect sizes. We did not meet this sample size goal, due to the difficulty in catching grackles and the large time commitment for serial reversal learning, and so it is possible we failed to detect some relationships. However, the power analysis included many more predictor variables than we ended up using (see Changes after the study began) and was conducted before we determined that the serial reversal learning trained grackles to be significantly more behaviorally flexible than control grackles (Logan et al., 2023). Thus, the increased difference in flexibility between control and trained grackles, also reflected in the ϕ and λ values, should increase our power to detect a relationship between these behavioral traits and flexibility, if it exists. Nevertheless, future research should evaluate these relationships with larger sample sizes.”

Comment 9:
I just felt like overall there was very little attention paid to the differences between the experimental and control birds? Like the authors report the overall mean level differences in their behavioral measures, but it’d be really nice to see the figures color coded as well to indicate which birds were in which group. Along these lines I also couldn’t tell whether the other behavioral measures (boldness, exploration) were conducted before or after the manipulation? This seems really important for interpretation?

Response 9: All behavioral traits (boldness, exploration, persistence, motor diversity) were measured after the manipulation occurred. This was originally stated at the end of the Methods > Behavioral flexibility section (Lines 216-218 of the previous version, Lines 236-239 of revision 1):

“All measures of the behavioral traits exploration, boldness, motor diversity, and persistence were collected after the serial reversal learning training was complete. By experimentally increasing the difference in flexibility performance between trained and control grackles, we increased our ability to detect a relationship, if it exists, between this trait and the other traits under investigation in this study.”

We now also state this earlier in Lines 191-199 of a new section in Methods > Test battery:

“...for all grackles we first assayed flexibility and implemented a flexibility training where half of the grackles underwent serial reversal learning and the other half received only one reversal and then control trials, described below. The training resulted in grackles more quickly changing

their behavior when reward contingencies changed, relative to control grackles (Logan et al., 2023). By experimentally increasing the difference in behavioral flexibility between control and trained grackles, we increased our power to detect relationships between flexibility and other traits. After grackles passed the behavioral flexibility training, they received the subsequent behavioral trait assays in a randomized order. Grackles were assayed twice for exploration and boldness, and given sessions with the MABs until they passed criterion. Because there were two MABs, we also have two measures of persistence and motor diversity for each individual."

We modified Figures 2, 3 and 4 to distinguish data points from individuals in the trained vs control groups.

Comment 10:
Anonymous reviewer 2

This study tested:

1- if exploration/boldness/persistence/motor diversity are repeatable traits in their study system.

2- if behavioural flexibility, measured as reversal learning, was related to exploration/boldness/persistence/motor diversity in wild caught great-tailed grackles.

3- (not introduced in the introduction but can be found in the methods and results sections) if exploration and boldness are repeatable in the wild vs in captivity

I would have a few main comments that would need to be addressed before publication, in my opinion:

I would suggest to keep points 1/2 in this paper and remove point 3. The goals are different and do not really fit in the same paper. Goal 3 is actually not present in the introduction of the present paper, but we can find the methods and results in the present paper. Moreover, goal 3 does not really have results as the sample size is too small.

Response 10: Good point! We moved the information about the captive vs wild hypothesis to the Supplementary materials 2 section. We also now mention in the Changes after the study began section that we attempted to measure wild individuals but could not get a large enough sample, Lines 422-425:

"We preregistered that we would compare performance on the boldness and exploration assays between grackles in the aviaries and those tested in the wild. However, we were unable to collect a large enough sample size to quantitatively test this hypothesis, therefore we present what we have in Supplementary Material 2."

Comment 11:
Is this study using the same data as Logan et al. 2016b? Because the question - if reversal learning is linked to behavioural traits like exploration persistence and motor diversity - seemed to be explored in both this paper and the present study? The differences would need to be emphasized to better understand the aim of the present study.

Response 11: Thank you for pointing out that this was not clear. This study uses different data than the Logan et al. 2016b study. The biggest difference between that older study and the present study is that they did not conduct a serial reversal learning training experiment to increase grackle behavioral flexibility. All of the apparatuses and analyses used in the 2016 study are different from the present study. In essence, the 2016 study provided a baseline upon which we improved to better understand the mechanisms of behavioral flexibility and how, or whether, it relates to other behavioral traits.

We added this information in the introduction to Lines 90-92: “Whereas, an earlier study in great-tailed grackles using different behavioral assays found no relationship between flexibility and any other behavioral traits, including persistence and motor diversity (Logan, 2016a).”

And we also added more details in Lines 118-123: “In a previous study with a smaller sample size (Logan, 2016a), we found no evidence for significant correlations between flexibility and the behavioral traits exploration, boldness, persistence, and motor diversity. However, this result could stem from the small sample size and lack of power to detect a relationship with a small effect size, or methods that do not result in repeatable performance. Based on this preliminary evidence, in the present study we increased our power to detect a relationship by training some individuals to be more flexible before measuring the other behavioral traits.”

Comment 12:
It is not clear to me, the purpose of training grackles for higher behavioural flexibility in the case of this study (L217-218)? These are potentially correlated traits that can be linked to behavioural flexibility, but there is no causation hypothesis between them. What it the goal? And the hypotheses? I do not understand how increasing behavioural flexibility on a task will have something to do with the ability to detect a relationship with other traits measured using other tasks? This might need more explanation here.

Response 12: While there is natural variation in how flexible grackles are, the training increased this variation such that grackles that were initially slow in their first reversal increased their switching speed to match the fastest grackles. If behavioral flexibility influences or is influenced by other traits like exploration, boldness, persistence or motor diversity, then we would expect training that increases behavioral flexibility to also affect performance on tasks measuring those other traits. By pushing the flexibility performance in trained grackles to the highest naturally observed, we expect a larger difference between trained and control grackles in behavioral trait performance (if a relationship between these traits exists) leading to a larger, more detectable effect size.

In a separate study, we found that the flexibility trained grackles were better at problem-solving (Logan et al. 2023) and showed more diverse foraging behaviors in the wild (Logan et al., 2025). This indicates that the flexibility training does impact behavior in contexts outside of reversal learning. Consequently, we are using this training to investigate whether flexibility is linked with these other behavioral traits to better understand the structure of behavioral mechanisms and validate methods for quantifying behavior (Logan et al. 2022).

We added more details about this in the Introduction in Lines 122-123: “in the present study we increased our power to detect a relationship by training some individuals to be more flexible before measuring the other behavioral traits.”

We also described the training more in a new paragraph in the Methods called “Test battery”, Lines 191-196:

“...for all grackles we first assayed flexibility and implemented a flexibility training where half of the grackles underwent serial reversal learning and the other half received only one reversal and then control trials, described below. The training resulted in grackles more quickly changing their behavior when reward contingencies changed, relative to control grackles (Logan et al., 2023). By experimentally increasing the difference in behavioral flexibility between control and trained grackles, we increased our power to detect relationships between flexibility and other traits.”

Comment 13:
L53: Behavioural flexibility, measured as reversal learning performance, is not rarely tested at the individual level. What is rarely tested is its link with actual individual traits that would indicate greater flexibility (e.g., novel food type, habitat range, exploration etc.).

Response 13: We modified this sentence (now in Line 54-55) as you suggest to read:

“The relationship between behavioral flexibility and adaptation to anthropogenic change is rarely directly tested at the individual level”.

Comment 14:

L57-60: I would not cite a preprint not yet reviewed as evidence (Logan et al. 2024) – or maybe it has been reviewed ??

Response 14: We understand and agree with the hesitation about citing work that has not been reviewed. This citation is for a post-study manuscript that is currently in round 2 of review at PCI Ecology (the first round of review did not flag up any issues about the link between flexibility and foraging diversity). It was also thoroughly reviewed as a preregistration (reviews are openly available here https://ecology.peercommunityin.org/articles/rec?id=35) and received in principal acceptance at PCI Ecology.

We updated the citation to indicate that it is in review at PCI Ecology.

Comment 15:

Add 1-2 examples – maybe from other species – to this section to describe a little bit more what has been found on this question (L 60-61).

Response 15: We added one example from mountain chickadees to Lines 57-65 to address this comment:

"The few studies that have directly related environmental adaptation to behavioral flexibility through measures of reversal learning show that flexible behavior can be closely linked with the current environmental niche. For example, mountain chickadees that live in harsh, high elevation environments perform worse on reversal learning tasks relative to lower elevation, milder climate individuals [@croston2017predictably]. This suggests that individuals that have a wider range of food options and a reduced reliance on cached food in milder climates require more cognitive flexibility to switch between food types. Additionally, new evidence from great-tailed grackles showed that more flexible individuals also demonstrate greater foraging diversity in the wild [@logan2024flexfor], and were better able to innovate solutions on a novel foraging apparatus [@logan2023flexmanippcj]."

Comment 16:

L67-74: I would add definition to the behavioural traits cited and add hypotheses- what link is expected -for boldness persistence and motor diversity too. So that we understand why those traits are cited here.

I just read that it has been done in the next paragraph, maybe both paragraphs should be joined.

Response 16: Good point, we simplified some of the redundancy and combined these paragraphs into one in Lines 70-91 (changes in bold, below):

Although flexibility has been the trait that much research has focused on to understand how behavior can impact adaptation to anthropogenic environmental changes, individual differences in other traits like exploratory tendency, boldness, persistence, or motor diversity could also play a role and correlate with behavioral flexibility [@logan2016behavioral; @sol2002behavioural]. To distinguish whether observed behavior in the wild or performance on behavioral trait assays are motivated by one or more distinct traits, it is important to measure multiple traits in the same individuals [@carter2013animal]. **However,** evaluation of the relationship between flexibility and other behavioral traits has produced inconsistent results [@dougherty2018linking; @logan2016behaivoral]. In one well studied avian group, the Paridae, **flexibility is related to exploration, which increases the likelihood of encountering fitness-enhancing resources in novel environments [@canestrelli2016bolder; @griffin2016invading]. This might imply**

that they are not two distinct traits, but the direction of the relationship is inconsistent across species **[positive: @herborn2014personality, @rojas2020exploration; negative: @amy2012worms].** Individuals approaching a potentially threatening aspect of the environment require a certain degree of boldness [@mccune2018evidence]. However, the relationship between boldness and flexibility can be positive [@titulaer2012personality], negative [@bensky2022behavioral; @bebus2016associative], or neutral [@guenther2014learning; @de2022bold]. Theoretically, persistence should inhibit flexibility because it results in perseverating on a previously rewarded behavior rather than changing to a more productive behavior for a given circumstance [@morand2022cognitive]. In contrast to persistence, motor diversity is theoretically positively correlated with flexibility because it implies that the individual has a repertoire of different behaviors it is able to choose from to match each circumstance [@diquelou2016role]. Research in squirrels supports this prediction [@chow2016practice], where the more flexible individuals were less persistent and more likely to use diverse motor behaviors. Whereas, an earlier study in great-tailed grackles using different behavioral assays found no relationship between flexibility and any other behavioral traits, including persistence and motor diversity [@logan2016behavioral].

Comment 17:

L75: I don't think the examples cited are 'experimental' evaluation, an experimental test would need a manipulation of a trait to test its effect on another. I think the authors cited correlative studies.

Response 17: See Response 16, above for changes to this paragraph that address this comment as well.

Comment 18:

L77-78: this ref refers to one species and thus does not illustrate inconsistencies, maybe it should be more described (like for exploration L80-83).

Response 18: Apologies, this was a typo. See Response 16 for the additional citations we meant to include here that show positive and negative correlations between flexibility and exploration.

Comment 19:

L78-79: same comment as above. I don't think this sentence is needed as concrete examples are provided after.

Response 19: We deleted this sentence.

Comment 20:

L107-108: I do not understand this sentence.

Response 20: We edited this sentence to hopefully be more clear, now in Lines 108-110:

“If behavioral traits are heritable, multiple traits can become linked through natural selection such that individuals that show high values on one trait (e.g, behavioral flexibility), will consistently display high values on a linked trait (e.g., exploration).”

Comment 21:

It is written that methods have been already reviewed elsewhere, so I guess they are relevant, but if this paper is to be published by itself, I would re work this part to be more clear. It was hard to follow, especially the behavioural flexibility section: I did not understand what test was done and why, how variables were measured etc.

Response 21: We added more details describing the method we used to measure behavioral flexibility, reversal learning. In Lines 201-222 we now include the following text:

“We used the reversal learning paradigm to measure flexibility as the ability to change behavior when circumstances change. In the first phase of reversal learning, subjects learn an initial association between a stimulus (here, color) and food. The reversal phase then occurs where the food is switched to the other color and the measure of flexibility is how quickly the subject learns the new food-color association. The methods for the initial association and the reversal trials are identical, where, on each trial, grackles could choose to look inside one of two colored containers for food (Fig. 1a). After they make a choice, the experimenter removes both containers, refills the food if necessary, then replaces the containers for the next trial. The side that the rewarded container was on was pseudorandomized to never be on the same side more than twice in a row to inhibit grackles from forming a side bias. When grackles showed a significant preference for the rewarded color in the initial association phase, demonstrated by choosing correctly on 17 out of the most recent 20 trials, we switched the location of the food to the other color container (a “reversal”). We measured baseline flexibility as the number of trials it took grackles to choose correctly on 17 out of the most recent 20 trials in this first reversal to demonstrate a change in preference to the second colored container. The flexibility training consisted of a randomized subset of grackles (n = 8) that received serial reversals where we switched the location of the food in multiple reversals after the grackle passed criterion in each reversal. Serial reversals continued until grackles were switching their preference in each reversal quickly enough to meet our experiment’s passing criterion of two consecutive reversals in 50 trials or fewer. We chose a criterion of 50 trials based on an earlier study of grackle reversal learning performance [@logan2016behavioral] where 50 represented an approximately 30% increase in the speed that grackles switched their preference in the first reversal [@logan2023flexmanippcj]. Grackles needed 6-8 reversals to pass this serial reversal training.

Instead of serial reversals, control grackles (n = 11) received equal testing experience with two identically colored containers, both containing a food item."

Comment 22:

I would add more details, such as:

L176: when where and how were the birds caught? Catching methods can have impacts on the personality profiles of the individuals caught.

Response 22: Good point. We added this information to the Methods in a paragraph called "Subjects" in Lines 169-173:

"Grackles were caught in the wild in Tempe, Arizona USA using mist nets, walk-in traps and bow nets. Trapping could occur at any time of day where grackles were active. While some trapping methods can select for subjects with certain traits (e.g., boldness: Biro & Dingemanse, 2009; but see Brehm & Mortelliti, 2018), mist nets are not visible to birds and no habituation is required, decreasing the probability of a selection bias for individuals that are more bold, food motivated, etc."

Comment 23:

L184-185: when and where?

Response 23: In response to your first comment regarding removal of this captive vs wild hypothesis, we deleted this sentence.

Comment 24:

L187: remove "as part of a different investigation", we need anyway this information here.

Response 24: You are right, we removed this phrase and added the necessary information about the behavioral flexibility methods. See Response 21.

Comment 25:

L188-190: I think it is important to add here the full description of the methods even if it has been described elsewhere. Some details, such as a validation test or a figure, can be cited from another paper but I think it is important to report the full methods for the study itself. For example, what is the training protocol?

Response 25: See Response 21.

Comment 26:

L192-193: “to switch their preference in the first reversal and search primarily in the second color container” the measure is not clear to me, is it the number of trials it took to 17/20 correct trials during the reversal?

Response 26: You are correct that behavioral flexibility is measured as the number of trials until grackles chose correctly on 17 out of the most recent 20 trials. See Response 21 for how we clarified this in the manuscript.

Comment 27:

L194-198: I am sorry, I do not see the point of this for this study?

Response 27: We added more details about the behavioral flexibility training (see Response 21) to address your comment. Also see Response 12.

Comment 28:

L196: why 50 is fast? What are the numbers that support this choice?

Response 28: We chose this criterion based on data on performance of grackles in their first reversal in the Logan et al. 2016 study. The individuals included in this previous research passed their first reversal in 70-130 trials and therefore a criterion of 50 or fewer trials is a significant increase in this latency to switch their preference. This criterion was validated in Logan et al. 2023 where we found that reversal learning performance of additional grackles (the flexibility data used here) demonstrated that 20% of grackles could pass the initial reversal in 50 or fewer trials. Thus, this criterion results in behavioral flexibility that is at the highest end of the natural variation in performance. We added the following to Lines 218-220 to clarify this:

“We chose a criterion of 50 trials based on an earlier study of grackle reversal learning performance (Logan, 2016a) where 50 represented an approximately 30% increase in the speed that grackles switched their preference in the first reversal (Logan et al., 2023).”

Comment 29:

L199-215: I do not understand this part, this was not in the aims of this study or I missed something? A MAB is mentioned for motor diversity assays, not for behavioural flexibility. This is really not clear. I would just describe the variable used to measure behavioural flexibility in this study (without referring to another).

Response 29: Thank you for the feedback that this was confusing. We removed this paragraph, added some of the details to the Changes after the study began section, and expanded on the subsequent Behavioral flexibility methods section to describe the variables actually used to measure behavioral flexibility in this study (changes in bold) in Lines 223-235:

"From the performance of each individual on reversal learning, we **used Bayesian reinforcement learning models** to create the Flexibility Comprehensive variables by modeling all of the choices that individuals made during the serial reversal learning experiment, and the uncertainty around these choices. **Because we include the sequence of all right and wrong choices individuals made during reversal learning, these variables more effectively represent flexibility compared to more commonly used variables such as the number of trials to reverse a preference. The details of this model and the validation of it as a measure of flexibility are described elsewhere (Blaisdell et al., 2021; Lukas et al., 2022). The Flexibility Comprehensive variables consist of** two components: ϕ (the Greek letter phi) as the rate of learning to be attracted to a color option and λ (the Greek letter lambda) as the rate of deviating from learned attractions that were previously rewarded. **Thus, our two measures of behavioral flexibility, that we subsequently included as covariates explaining behavioral trait performance, were the Flexibility Comprehensive continuous variables or the dichotomous variable describing whether the grackle was in the flexibility trained or control group. There was one measure per individual for each of these variables."**

Comment 30:

L228-229/2039-240: I would move here the info about the variable chosen.

Response 30: No problem, we moved this information to the end of the Boldness and Exploration methods sections (now Lines 249-251 for Boldness and 261-265 for Exploration).

Comment 31:

L242-251: it is not clear how motor diversity and persistence are measured from the results of various tests?? Means are used? Or number of trials are added?

Response 31: For both measures we used *a sum* of the number from each different test. For motor diversity, we counted the number of different motor actions (based on the ethogram) that an individual demonstrated on each of the two multiaccess boxes. To quantify repeatability of motor diversity, we analyzed whether the number of motor actions on the two MABs were consistent within individuals across these two contexts. Similarly for persistence, we summed the number of touches a grackle made to each different test, including novel environment, novel object, boldness hawk, cat and pigeon, as well as the two MABs. We then quantified repeatability of persistence to test whether the number of touches a grackle made to each test apparatus was consistent across these different items.

Motor diversity was not repeatable, so we did not relate it to flexibility. Persistence was repeatable (but see Response 41), and so in the model that evaluated persistence as a function of flexibility we used *the average value* for number of (functional) touches across the different test apparatuses as the dependent variable.

We clarified this in Lines 268-274:

"For each grackle, we summed the number of distinct motor actions they used while interacting with each MAB, resulting in two values for each grackle. We quantified persistence as the number of touches to a novel apparatus per trial time [@logan2016behavioral; @griffin2015innovative], where the novel apparatuses included the novel environment and novel object from the exploration assays, the potentially threatening boldness objects, as well as the two MABs. We summed the number of touches grackles made to each apparatus, resulting in a value of persistence for each test apparatus, if the grackle received that test (e.g., two grackles did not participate in the MAB tests). We further distinguished touches to the MABs based on whether they were functional (touches to the doors or loci that could result in getting the food item) or nonfunctional (touches to the side of the box that would never result in food)."

Comment 32:

L281-288: I have to say that I have no clue about what has been done here, but maybe a ref or two would be useful to show that this is the usual way to do it?

Response 32: No problem, we added a citation to this paper here, Line 305:

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935-956.

Comment 33:

L290: so which measure, the first one or the second one or a mean?

Response 33: See the above Response 31

Comment 34:

L293-295: sex should be included too.

Response 34: We had not preregistered that we would look for an effect of sex on performance because we had previously found no sex difference in behavioral flexibility. However, as behavioral flexibility is no longer the dependent variable, sex might influence performance on the behavioral trait tasks. We re-ran all of our models including a covariate for both Sex and Age (in response to your Comment 36). All results were qualitatively identical, indicating these

variables did not affect performance on the behavioral trait tests. To avoid overfitting our models with our sample size, we do not include the Sex and Age covariates in the models we present in the manuscript. However, we add these details to the “Changes after the study began” section in Lines 402-409:

“We preregistered that we would include “Age” as a covariate in our models relating performance on the behavioral trait assays to flexibility, if we tested juveniles as well as adults, though our plan was to only test adults. Our sample ultimately included two juveniles because the grackles were more difficult to catch than expected and we struggled to meet our minimum sample size. Similarly, it is possible that Sex could influence performance, but we only tested 4 females because they were more difficult to trap than males. We did not find that including a covariate for Age and Sex changed any of our results (repeatability or relationship with flexibility). Therefore, to maintain greater statistical power, we decided to not include Age or Sex as covariates in the final models.”

Comment 35:

L370-372: but then the number of models run will increase? I would keep Behavioural flebility as a response variable, and all behavioural measures as independent variables, but separate in two analyses, one for the group with high flexibility, and on for the control group. That way you will see by comparing the results and the estimates if the relationships are stronger?

Response 35: We referred back to Hernán & Robins (2006) to help us get clear about this, and they state that it is crucial that the behavior that was manipulated is not the outcome variable. The assignment to a condition (trained or control) was the instrumental variable (Z) in a randomized experiment, where individuals experienced a treatment (X; serial reversals or only 1 reversal) with behavioral trait performance (exploration, boldness, persistence, motor diversity) as the outcome variable (Y).

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. Journal of Epidemiology & Community Health, 60(7), 578-586.

Comment 36:

L381-386: I guess a good way to test that it is ok to do so is to compare the results with and without the 2 juveniles and see if they are qualitatively the same? I would do the same for sex.

Response 36: Good point, we did this for Age as well as Sex (see also Response 34, above). These covariates did not have any significant effect, so we ultimately dropped them to avoid overfitting models with our sample size. We added this information in Lines 288-291:

“Due to our unbalanced sample of sex and age we checked whether these variables significantly impacted the response. We found that these covariates did not have a significant

effect on any of the models (described below), so we omitted them from the final models (see *Changes after the study began* section)."

Also Lines 402-409:

"We preregistered that we would include "Age" as a covariate in our models relating performance on the behavioral trait assays to flexibility, if we tested juveniles as well as adults, though our plan was to only test adults. Our sample ultimately included two juveniles because the grackles were more difficult to catch than expected and we struggled to meet our minimum sample size. Similarly, it is possible that Sex could influence performance, but we only tested 4 females because they were more difficult to trap than males. We did not find that including a covariate for Age and Sex changed any of our results (repeatability or relationship with flexibility). Therefore, to maintain greater statistical power, we decided to not include Age or Sex as covariates in the final models."

Comment 37:

L420: data should enable to confirm this: did more grackles participated/stayed longer on the ground in the second test than in the first time, on average (not clear from Fig2)? One possibility is also that the objects were too scary, leading to a floor effect.

Response 37: Yes, grackles on average did stay longer on the ground in the second trial for the threatening conditions. We added this additional analysis to this section in Lines 451-454 to clarify:

"We conducted an *unregistered analysis* and found that grackles did spend significantly longer on the ground during the cat, hawk and novel object (which the grackles considered threatening, see below) trials, relative to the first trials (Poisson model: $*ß* = 0.85$, $p < 0.01$)."

Comment 38:

L462-463: is it possible that the repeatability effect is found because lots of results were just 0 peck? I think persistence measured during boldness or exploration tests would not indicate the same 'persistence' as during problem-solving tests (this is partly confirmed with the figure 4 where nearly all results from bolness and exploration tests are around 0), I would only use persistence during problem-solving.

Response 38: Repeatability focuses on differences among individuals relative to differences within individuals on repeated assessments. When all grackles are similarly unwilling to touch the boldness or exploration apparatuses, then these zeros do not significantly affect the repeatability. Indeed, when we model repeatability of the number of touches to only the multiaccess boxes (excluding boldness and exploration) we get close to the same value for repeatability and it is still statistically significant.

We think it is important to show the variation in how much grackles interacted with each task and that those that interacted more with the MABs also interacted more with the boldness and exploration tests, even if it was a small number of touches per time. It is impossible to know whether grackles perceived the boldness and exploration apparatuses the same as the MABs. However, these repeatability results indicate that touches were consistent and may reflect a similar kind of persistence across MABs and boldness/exploration.

Comment 39:

L482-484: the repeatability test has been done only on latencies, so I would use only this measure (if I remember well, it was written earlier in the MS that analyses with flexibility would be done only on repeatable measures). And then, which measure has been used for this analysis, the first the second or a mean?

Response 39: Both the “latency to approach” and “duration near” variables were repeatable in the exploration assay. This is stated in the Results > Repeatability > Exploration section, now in Lines 478-480. We think this might have been confusing because the plot for the exploration results showed only the “latency to approach” variable. We updated this figure (Figure 3) to include both the variables we used to measure exploration.

We used all of the data in the exploration analysis, so there were repeated measures of each exploration variable. To account for this, we included a random effect for individual ID. In the persistence models, we used the average number of touches to improve the fit of the models. This is explained in the “Changes after study began” section, Lines 415-418:

“We preregistered that we would use all of the data, including the repeated measures, with a random effect for individual ID in a Poisson model. However, the full data set was zero-inflated. Because persistence was repeatable across assays, we took the average for each individual to use as the dependent variable in our model.”

Comment 40:

L492: ‘in contrast’ with what?

Response 40: Thank you for pointing out the lack of logical flow here. We removed this phrase.

Comment 41:

L493: as above, the repeatability test has been done only on the total number of pecks. And here only persistence measures with MAB are used?

Response 41: We see how this was unclear. We preregistered predictions that flexibility may relate to persistence differently based on whether the touches were functional or nonfunctional.

All touches to the boldness and exploration apparatuses are considered functional because in these tasks there is no type of touch where grackles can get a food item, but they potentially learn something about the novel item through each touch. Similarly, on the two different multiaccess boxes (MABs), grackles can make functional touches which result in a food item and the grackle will learn something about the MAB as a result of those functional touches. However, grackles could make touches on the MABs that would never result in a food item and so were considered nonfunctional.

For this analysis, we preregistered that we would model the relationship between flexibility and persistence as the number of functional touches per time to all apparatuses (which were repeatable). We preregistered that we would also conduct a separate model for the relationship between flexibility and nonfunctional touches (which only occur on the MABs). As a result of your comment, we now realize that we did not test whether nonfunctional touches to the two different MABs are repeatable on their own before assessing the relationship with flexibility. We added this analysis and found that nonfunctional touches to the 2 MABs are not repeatable. Therefore, we removed the analyses relating nonfunctional touches to behavioral flexibility.

We added this analysis to “Changes after the study began” in Lines 410-413.

“We added an additional persistence repeatability analysis to test whether nonfunctional touches were consistent across the two different MABs. We preregistered that we would separately evaluate the relationship between flexibility and functional or nonfunctional touches, but, because flexibility was originally the dependent variable, we did not preregister this repeatability analysis.”

Note that removing the nonfunctional touch analysis resulted in the removal of the result that grackles with higher phi and lower lambda were more persistent with nonfunctional touches, such that now there are no relationships between flexibility and persistence.

Comment 42:

I think the discussion will have to change a bit. Negative results are discussed like real negative results but it should be emphasized a little bit more that they could also just reflect a problem with the methods or a lack of statistical power.

Response 42: We added to the discussion a paragraph regarding the influence of our sample size on our power to detect relationships between the repeatable traits and behavioral flexibility (see Response 8, above). We also deleted the paragraph discussing the significant relationship between persistence and flexibility that went away when we determined that nonfunctional touches were not repeatable.